DIBELS Next Development: Findings from Beta 2 Validation Study

Elizabeth N. Dewey
Rachael J. Latimer
Ruth A. Kaminski
Roland H. Good

Dynamic Measurement Group

Technical Report 10

Suggested Citation

Dewey, E. N., Latimer, R. J., Kaminski, R.A., & Good, R.H. (2012). *DIBELS Next Development: Findings from Beta 2 Validation Study* (Technical Report No. 10). Eugene, OR: Dynamic Measurement Group.

Author Note

The authors thank the faculty, staff, students, and parents of participating schools for their effort and support during the course of this study. Correspondence regarding this manuscript should be addressed to Elizabeth N. Dewey, Dynamic Measurement Group, 132 E. Broadway, Suite 636 Eugene, OR 97401; Email: edewey@dibels.org.

Acknowledgements

This technical report has been a collaborative effort of many. The talents and efforts of hundreds of individuals contributed to the successful completion of this report, including Dynamic Measurement Group research scientists and staff, educators and school personnel, children and parents, and publishing partners. We deeply appreciate the contributions of Kelli D. Cummings, Ph.D, who served as the Principal Investigator of this study. We also extend our thanks to the following people for their contributions to this technical report.

Research Support
Joshua Wallin
Annie Hommel
Douglas P. Rice
Maya E. O’Neil, Ph.D.
Chantal Dufour-Martel, Ph.D.
Kelly A. Powell Smith Ph.D.

Field-Site Coordinators
Carol Trumbo

Analysis of DIBELS Next: Findings from Beta 2 Validation Study

The purpose of this technical report is to detail the findings from a validation study evaluating the measures that comprise DIBELS *Next*. In this report we will provide information about the reliability of two measures new to DIBELS Next: *First Sound Fluency* and *Daze.* We will also evaluate the effects of changes to the directions developed for DIBELS Next for *Letter Naming Fluency*, *Phoneme Segmentation Fluency*, and *Nonsense Word Fluency*. Finally, the individual scores of *Nonsense Word Fluency* will be evaluated. Descriptive statistics and correlational data will provide the foundation of this review. For further information regarding the development of these measures, as well as additional technical reports, please visit: www.dibels.org .

Method

Participants. The participants in this study were students in kindergarten through fifth grade in five elementary schools from one school district located in the Pacific Northwest region of the United States. The school district was one participant out of thirteen school districts that were involved in a larger study during the 2008-2009 school year on DIBELS Next measures.

The participating school district was recruited from a list of sites that had previously volunteered to participate in DIBELS-related research, and was chosen because of its proximity to the research organization. All students at the participating schools who were typically assessed as part of their school's universal screening activities were included in the study.

Data for this study were captured, primarily, through an extant database (i.e., the DIBELS Data System, (DDS), https://dibels.uoregon.edu/). As a result, some demographic information was not reported at the student level (i.e., data for demographic categories such as free and reduced lunch eligibility and special education eligibility). All five schools that elected to participate in this study reported data for the race and ethnicity category, and additional demographic data was gathered at the school level from the National Center for Education Statistics (NCES) website (http://nces.ed.gov/) for the 2007-2008 school year. Table 1 summarizes these data. District data is the aggregate of participating school level data reported by NCES. This may include data for students in grades not included in the study (i.e., 6th, 7th and 8th) and may not include data from those that elected not to report.

Table 1

Beta 2 Validation Study School District Demographics

Census Region, Division	West, Pacific
State	OR
DIBELS® Experience	2 years
District-Wide Information	
Total Schools	9
NCES Surveyed Total	1518
ELL Students	60
Students with IEP's	490
Expenditure Per Student	$8,894
Race/Ethnicity	
American Indian or Alaska Native	254 (17%)
Asian	19 (1%)
Black or African American	18 (1%)
Hispanic	126 (8%)
White	1077 (72%)

NCES Surveyed Total	1494
Beta 2 Validation Schools	
Total Schools	5
Total Teachers	72
Student:Teacher Ratio	21.1
Total Students that Qualify for Free Lunch	658 (43%)
Total Students that Qualify for Reduced Lunch	227 (15%)

Source: U.S. Dept. of Education, National Center for Education Statistics, Common Core of Data (CCD) for the 2007-08 school year. Data is based on actual reported numbers and may not include students who elected to not report these data. "District Wide" data includes schools and grades not involved in this study. "Beta 2 Validation Schools" data is the sum of school-level data reported to NCES and may include grades, such as 7, 8, that were not included in the study. Percent of population is indicated in parentheses.

Demographic information-student level. To further explain the characteristics of the sample, we include information about the benchmark status of students across the three data collection time points. The initial skill level of students in the fall is actually the strongest determinant of gains during the year (Hedges & Hedberg, 2007; Bloom, Richburg-Hayes, and Black, 2007; and Slavin, 2008) and the best way to at least partially equate students and schools. No other predictors that are routinely collected, especially demographic information, are as good at explaining both student and school level differences (Stoolmiller, et al., 2008). Table 5 includes percentages of students at each benchmark level. The typical percentage of students at benchmark is around 60% in practice, thus, it appears that our sample is below average overall, with student performance above average in kindergarten (e.g., PSF at all times of year) and below average on most other measures in all grades (e.g., DORF in all grades). For further information on the benchmark goals and cut points for all DIBELS measures, please see the *DIBELS Next Assessment Manual* (Good III, R. H., Kaminski, R. L., 2010).

Procedures

The data for this study were collected by trained staff from participating school districts in accordance with the site's existing procedures. District coordinators were trained on all new measures (see *Measures* section) by a Dynamic Measurement Group (DMG) Research Scientist via webcast prior to beginning-of-year data collection. The webcast was four hours in duration, and included ample time for practice with simulated assessment activities. All district coordinators were responsible for checking the members of their team for reliability of test administration.

The participating site had access to the Beta release of these measures from a password-protected download site. The site was instructed to assess all students at their assigned grade level for benchmarking, three times per year, and also had access to progress monitoring materials that they could use in accordance with their current practices. For the purpose of this study, only benchmark data were shared and analyzed. Table 2 presents the measures administered for the study by grade and time of year.

Table 2

Table of Timelines for Measure Administration

	Kindergarten			First grade			Second grade		
Measure	Beg.	Mid.	End	Beg.	Mid.	End	Beg.	Mid.	End
FSF	X	X							
LNF	X	X	X	X					
PSF		X	X	X					

NWF		X	X	X	X	X	X		
DORF					X	X	X	X	X
RTF					X	X	X	X	X

	Third grade			Fourth grade			Fifth grade		
Measure	Beg.	Mid.	End	Beg.	Mid.	End	Beg.	Mid.	End
DORF	X	X	X	X	X	X	X	X	X
RTF	X	X	X	X	X	X	X	X	X
Daze	X	X	X	X	X	X	X	X	X

Note. Beg. = Beginning of year; Mid. = Middle of year; End = End of year; FSF = First Sound Fluency; LNF = Letter Naming Fluency; PSF = Phoneme Segmentation Fluency; NWF = Nonsense Word Fluency; DORF = DIBELS Oral Reading Fluency; RTF = Retell Fluency; Daze = DIBELS Maze. At the time of this study, Retell Fluency remained an optional DIBELS measure, and as a result, schools only administered this measure if they self-selected to do so. Alternate forms of all middle-of-year DIBELS measures were administered two weeks after middle-of-year benchmark assessment.

DIBELS Measures[1]

First Sound Fluency (FSF). Formerly called *First Sounds* (see Kaminski, Baker, & Smith, 2006), FSF measures the ability to isolate the first sound in a word, which is an important phonemic awareness skill that is highly related to reading acquisition and reading achievement (Yopp, 1988). FSF is used as a measure of developing phonemic awareness at the beginning and middle of kindergarten. The assessor says a series of words one at a time to the student and asks the student to say the first sound in the word. On the scoring page, the assessor circles the corresponding sound or group of sounds the student says. Students receive 2 points for saying the initial phoneme of a word (e.g., saying the /s/ sound as the first sound in the word *street*) and 1 point for saying the initial consonant blend, consonant plus vowel, or consonant blend plus

[1] A complete description of the DIBELS Next Measures is available in the DIBELS Next Administration and Scoring Guide.

vowel (e.g., /st/, /str/, or /strea/ for *street*). A response is scored as correct as long as the student provides any of the correct responses listed for the word. The total score is based on the number of correct 1- and 2-point responses the student says in 1 minute.

Items for all FSF probes were selected from a word pool consisting of single-syllable words. Initial work on this word pool was derived from a study of preschool measures of early literacy (Kaminski, Baker, Chard, Clarke, & Smith, 2006). Words were excluded if they were deemed inappropriate (e.g., rob, knife) or if they began with the initial phonemes /b/, /d/, /p/ or /g/, followed by the /u/ sound (e.g. duck), as such words cannot be scored differentially due to confusion with the schwa sound. The final word pool consisted of 871 words, 3 of which were used as example items and so do not appear as test items.

Half of all test items were categorized as initial blends, and half were categorized as initial phonemes. All probes used in the study consisted of 30 items so that, within each group of four items on the probes, two words began with blends, one began with an initial phoneme with a stop sound, and one began with an initial phoneme with a continuous sound. The order of the categories within these groups of four was randomly determined, except for some easier categories placed at the beginning of each probe. The two remaining items on each probe were an initial phoneme with a stop sound and a blend. Probes were stratified so that the same difficulty levels appeared in the same locations on each probe.

Letter Naming Fluency (LNF). LNF is a brief, direct measure of a student's fluency naming letters, and assesses a student's ability to recognize individual letters and say their letter names. Fluency in naming letters is a strong and robust predictor of later reading achievement (Adams, 1990) but is not a powerful instructional target, i.e., focusing instruction on letter names has not been shown to lead to better reading outcomes. The student is presented with an 8.5" by

11" sheet of paper with randomly ordered upper- and lower-case letters, and asked to verbally provide the names of the letters. The student is allowed one minute, and the final score is the number of correct letter names produced in that minute.

All upper- and lower-case letters in the English alphabet are used. The 26 upper-case and 26 lower-case letters are divided into three categories based on relative difficulty, with 18 letters in the easy category and 17 letters each in the medium and hard categories. A randomly selected letter from the easy category was used as the first item, and then 17 triads were constructed, with a triad including one randomly selected letter from each category, easy, medium, and hard. The first triad was placed with the easy letter first, the medium letter second, and the hard letter third. The other 16 triads randomized the order of the difficulty categories within each triad. The process was then repeated, to include another set of 26 upper-case and 26 lower-case letters. The only difference in procedure for the second set of 52 letters was that the difficulty categories in the first triad were randomized. There are 104 total test items arranged in eleven rows. The first ten rows contain ten letters each. The eleventh row contains the final four letters followed by the first six letters that appeared on the probe, repeated in the same order.

Phoneme Segmentation Fluency (PSF). PSF is a test of phonological awareness (Kaminski & Good, 1996). The PSF measure assesses a student's ability to segment words into their individual phonemes fluently, and has been found to be a good predictor of later reading achievement (Kaminski & Good, 1996). The PSF task is administered by the examiner orally presenting words of two to five phonemes. It requires the student to verbally produce the individual phonemes for each word. For example, the examiner says "sat", and the student says "/s/ /a/ /t/" to receive three possible points for the word. After the student responds, the examiner presents the next word, and the number of correct phonemes produced in one minute determines the final score.

Items for all PSF probes were selected from a word pool consisting of 1,158 two- to five-phoneme, single-syllable words. The word pool was divided into four categories based on the number of phonemes, and difficulty factors such as blends and r-controlled vowels. Each probe contained 24 test items, including 16 from the easiest category, 6 from the next easiest category, and 1 each from the two harder categories. The order in which the categories appeared on the probes was randomly determined, except for some easier categories placed at the beginning of each probe. Probes were stratified so that the same difficulty levels appeared in the same locations on each probe.

Nonsense Word Fluency (NWF). The NWF task is a measure of the alphabetic principle—including both letter-sound correspondence and the ability to blend letters into words in which letters represent their most common sounds (Kaminski & Good, 1996). The student is presented with an 8.5” by 11” sheet of paper with randomly ordered VC and CVC nonsense words (e.g., sig, rav, ov) and asked to verbally produce the individual letter sound of each letter or read the whole nonsense word. For example, if the stimulus word is “sig” the student could say “/s/ /i/ /g/” or say the word “/sig/” to obtain a total of three correct letter sounds. The student is allowed one minute to produce as many letter-sounds as he or she can. There are two scores recorded for this measure: the total number of correct letter sounds (CLS) and the number of nonsense words read as whole words (WWR), with two additional scores representing the number of words read correctly (WRC) and the number of nonsense words sounded out and recoded (SOR). WRC is equivalent to the sum of SOR and WWR.

Items for all NWF probes were selected from a word pool consisting of 1,026 VC and CVC nonsense words, 2 of which were used as example items and so do not appear as test items. The word pool was divided into six categories based on the number of letters and whether the consonants in the nonsense word were easier or harder. Each probe contained 50 test items arranged in ten rows with five items each. Each row included two CVC words where both consonants were easy, one CVC word where the first consonant was easy, and one CVC word where the last consonant was easy. Each row also included one other item from one of the remaining categories (five rows with a VC word with an easy consonant, two rows with a VC word with a hard consonant, and three rows with a CVC word where both consonants were hard). The order in which the categories appeared in a row was randomly determined, except for the beginning of each probe in which items from easier categories were placed first. Probes were stratified so that the same difficulty levels appeared in the same locations on each probe. In addition, vowels were evenly distributed so that each row of five words included a single word with each vowel, a, e, i, o, and u.

DIBELS Oral Reading Fluency (DORF). DORF individually administered test of accuracy and fluency with connected text (Good & Kaminski, 2002). The student is presented with a reading passage on an 8.5" by 11" sheet of paper and is asked to do his or her best reading. The student is allowed one minute to read the passage, and the recorded scores are the number of words read correctly in that minute (DORF Words Correct) and the number of errors made. Calculated from these scores is the student's Accuracy score (DORF Words Correct divided by the sum of DORF Words Correct and the number of errors made). For standard benchmark assessments given at the beginning, middle, and end of the school year, the student is administered three passages. The median DORF Words Correct and Accuracy scores out of the three passages administered are recorded as the student's final score.

DORF passages were written according to specific criteria to ensure the appropriateness of the content. DORF includes a mix of different types of passages; approximately two-thirds of the passages in first through third grades were narrative and one-third were expository; one-third of the passages in fourth through sixth grades were narrative and two-thirds were expository. To prevent ceiling effects, the passage length in each grade is designed so that most students will not finish the passage in one minute.

The difficulty levels of the passages were targeted to grade-specific ranges using the DIBELS Readability Formula (Cummings, Wallin, Good, & Kaminski, 2007). Traditional readability formulas use indicators representing two of three aspects of passage difficulty that can readily be counted: (a) decoding difficulty (word length), (b) semantic difficulty (word frequency or rare words), or (c) syntactic difficulty (sentence length). In traditional readability formulas, the two aspects examined are combined into a single result, which means that the individual considerations are not examined in isolation many difficult words, for example, could be combined with short sentences to provide a misleading estimate of passage difficulty. The DMG Passage Difficulty Index combines indicators representing all three aspects, both in isolation and combined, to ensure that each consideration is within the target range for the grade level, as well as the overall composite of the three considerations. The indicators included in the formula are sentence length, word length, and percent of rare words.

Retell Fluency (RTF). Passage retell provides an indicator of reading comprehension. During retell, the student is asked to tell about what he/she has read. The assessor indicates the number of words in the retell that are related to the passage by drawing through a box of numbers. Following a hesitation of 3 seconds, students are prompted to tell as much as they can about the passage. If the student hesitates again for 5 seconds or longer, or if the student is clearly responding for 5 seconds in a way that is not relevant to the passage, the task is discontinued. The assessor must make a judgment about the relevance of the retell to the passage. Retell can be used from the middle of first grade through the spring of sixth grade.

Daze. Daze is the standardized DIBELS version of maze procedures for measuring reading comprehension (Good et. al, 2010). Daze assesses the student's ability to construct meaning from text using word recognition skills, background information and prior knowledge, familiarity with linguistic properties such as syntax and morphology, and cause and effect reasoning skills. Daze can be given to a whole class at the same time, to a small group of students, or individually. Using standardized directions, students are asked to read a passage silently and to circle their word choices. Approximately every seventh word in the Daze passages has been replaced by a box containing the correct word and two distracter words. The student receives credit for selecting the words that best fit the omitted words in the reading passage. The number of correct and incorrect responses are recorded. An adjusted score calculated by subtracting half the number of errors made from the number correct compensates for guessing.

Validation administration

***Reliability and Validity of DIBELS*.** Approximately two weeks after the middle-of-year benchmark assessment, six data collectors employed by DMG collected additional study data from students in all grades in all five schools within the district to evaluate the reliability and validity of DIBELS Next measures. This period of data collection is referred to as 'validation administration'. In four schools, alternate-form reliability data was collected on FSF. Alternate-form reliability data for Daze was exclusively collected at one school where no other alternate-form measures were given.

Change in Directions. In development for DIBELS Next, all measures went through an evaluation process where content, directions, and scoring procedures were examined. During the Beta 2 study, measures were administered at beginning-, middle-, and end-of-year benchmark assessment. For all benchmark assessments, DIBELS Next forms of LNF, PSF, NWF, DORF, and RTF were administered with DIBELS 6th Edition directions and scoring procedures. During validation administration, students were given alternate forms of these measures using directions and scoring procedures in development for DIBELS Next (referred to throughout this report as "DIBELS Next directions and scoring procedures").

Data Collection. Benchmark assessment data were collected by school-based personnel trained by DMG. Fidelity of assessment was evaluated following the midpoint of the study (i.e., middle-of-year benchmark administration). Site coordinators were asked to complete a “fidelity of assessment” checklist (contact the first author for details or to see a copy of the checklist). Data was retrieved from the DIBELS Data System.

Data on alternate forms and DIBELS Next directions were collected by DMG personnel. Testing took place outside the classroom at a desk in a room (music, library, gym, title, etc.) set aside for our use. All students were read an assent script before proceeding with the measures. Care was taken to engage students and develop an appropriate rapport. Total time away from the classroom varied between 5-15 minutes depending on grade level and individual student behavior. All students were awarded an incentive (sticker) once they completed the assessment.

A set of labels with student ID, School, Grade, and Teacher were created at DMG offices using the data files from the DIBELS Data System. The Beta 2 on-site coordinator for the participating district assisted DMG alternative-form data collectors throughout the data collection process by pulling students from their classroom and affixing the appropriate label to a booklet before a student was tested. In the few cases where no label was available for a student, the on-site coordinator hand wrote an ID number.

A total of 1,386 parent opt-out forms were distributed. A total of 41 students had parents that returned opt-out forms. A total of four students did not assent to continue participating in the study. Two students declined to continue mid-way through data collection. For Daze, a total of 18 students did not assent. Data on 543 students was collected on DIBELS grade-level booklet measures, and data on 146 students was collected on the Daze measure, from the 4 schools where non-Daze measures were administered. Data was collected on a total of 689 students.

All probes were scored immediately after administration except when a scoring question arose that warranted waiting until either the PI or a fellow examiner could offer clarification. Probes were rescored by different members of the original data collection team with no individual scoring the same probe twice. Data entry was completed at the research organization's office and reliability was attained through 100% redundant data entry. A total of six full booklets were spoiled during middle-of-year validation data collection due to DMG data collector error. A total of 19 individual measures were spoiled in otherwise valid booklets due to either DMG data collector error or interruption in administration.

Research Questions

The specific questions to be addressed by this study are:

1. What is the alternate-form reliability and the concurrent and predictive validity of DIBELS new measures *First Sound Fluency* and *Daze*?

2. What are the middle-of-year intercorrelational relationships and predictive validity to end-of-year DIBELS outcomes of DIBELS measures *Letter Naming Fluency*, *Phoneme-Segmentation Fluency*, and *Nonsense Word Fluency* with the modified directions and scoring procedures in development for DIBELS Next?
3. What is the effect, if any, of changes to directions and scoring procedures for DIBELS *Letter Naming Fluency*, *Phoneme-Segmentation Fluency*, and *Nonsense Word Fluency*?
4. How do DIBELS *Nonsense Word Fluency* scores *Words Read Completely and Correctly* (WRC), words *Sounded Out and Recoded* (SOR), and *Whole Words Read* (WWR) compare to each other and contribute to the NWF measure?

Results

Strict data management procedures were followed to assure that analyses were performed with accurate data. Data was collected at all three benchmark assessment periods plus the alternate-form assessment period, but this study will only evaluate data collected from middle-of-year benchmark assessment, middle-of-year alternate-form assessment, and end-of-year benchmark assessment. Data collected on second-grade DIBELS measures (DORF and RTF) was not assessed as part of this study, because DORF and RTF are evaluated only as they relate to NWF and Daze (both of which are not administered in middle-of-year and end-of-year second grade). Participant sample size was 609. Of these, 598 participants from kindergarten, first, third, fourth, and fifth grades had complete DIBELS data for all measures for all time points (middle-of-year benchmark assessment, middle-of-year alternate-form assessment, and end-of-year benchmark assessment).

We examined the data for scores that were invalid due to known errors with data collection (e.g., lack of fidelity to assessment procedures), invalid ranges (i.e., scores above the maximum possible for a given measure), or significant univariate or bivariate outliers. See measure-specific sections for details regarding the way in which these steps impacted the final sample sizes by measure, grade, and time of year.

Scoring Decision Rules for Evaluating Outliers. We examined all of the DIBELS data by measure, grade, and time of year for the presence of outliers and/or invalid scores. Almost every measure within each grade and time of year category had outliers that were more than 1.5 standard deviations above the mean, as is common in large datasets (Tabachnick & Fidell, 2007). We defined a severe outlier as scores that were *more than three standard deviations above the mean score* for that measure. We defined an invalid score as a score was greater than the maximum possible or the score was an illegal value when evaluated conjunction with another score (e.g., a combination of NWF scores of WWR = 5 and CLS = 7 is not a valid pair). Using these decision rules, no outliers were removed for this study, and invalid scores were set to missing. The following details the severe outliers and invalid scores discovered in the dataset evaluated for this study.

NWF. The maximum CLS score is 143, the maximum WWR is 50. One score for CLS, three scores for the alternate-form of CLS, two scores for WWR, and three scores for the alternate-form of WWR were classified as outliers in kindergarten. However, none of these scores were above the maximum allowable, so they were all included in further analyses including this measure.

DORF. The DORF WC and DORF Errors scores are limited by the number of words in the DORF passage, thus the maximum scores are pre-determined. There were no DORF Words Correct scores classified as outliers.

RTF. On the RTF measure, there is space provided in the scoring box for up to 94 words in the retell of the passage read in grades one through six. Though this does not technically limit the response by the student to 94 words (i.e., the examiner is not instructed to tell the student to "stop" at this point in the exam), it does *practically* limit the recorded score. Thus, it was determined that the maximum score would be limited to 94. Three Retell scores were classified as outliers, one in each of third, fourth, and fifth grade in addition to one score in second grade on an alternate-form probe. However, the outliers were not greater than 94, so all were included in further analyses including this measure.

Daze. The total number of words correctly selected (DAZ), the number of errors (DZE), and their sum are limited by the maximum possible score, which is limited by the maximum number of selection items on the probe which varies by passage and by grade. Two DZE scores were classified as outliers, one in third grade and one in fifth grade. Given that neither outlier was above the maximum possible, and the sum of correct and incorrect values did not exceed the maximum possible, these scores were allowed to remain in the data set and were included in additional analysis of this measure.

Descriptive Statistics and Quartiles for DIBELS measures

Letters have been assigned to the different administrations to make the relationships between measures and sets of directions clearer; time point A = middle-of-year benchmark administration given with DIBELS 6th Edition directions, time point B = middle-of-year validation administration given with directions in development for DIBELS Next (referred simply as DIBELS Next directions), and time point C = end-of-year benchmark administration given with DIBELS 6th Edition directions. At all three administrations, DIBELS Next assessment content (forms and booklets) were used. Only middle-of-year and end-of-year measures are examined.

Data is presented on DIBELS Next measures: FSF, LNF, PSF, NWF-CLS, NWF-WRC, DORF WC, RTF, and Daze adjusted score. Descriptive statistics are reported in Tables 3 and 4. Table 5 reports the percentage of students in each measure support category.

Table 3

Descriptive Statistics for the Beta 2 Validation Study Sample

		Middle of Year (Time Point A)		Middle of Year Validation (Time Point B)		End of Year (Time Point C)	
Measure	N	*Mean*	*SD*	*Mean*	*SD*	*Mean*	*SD*
				Kindergarten			
FSF	97	30.10	14.74	28.66	14.32	--	--
LNF	94	23.63	15.45	23.20	16.61	34.14	16.57
PSF	96	33.94	17.41	29.02	14.92	47.15	17.12
NWF-CLS	90	19.42	12.50	17.66	11.93	32.66	20.04
NWF-WRC	86	2.67	4.57	2.84	3.89	5.24	7.40
				First Grade			
PSF	71	52.99	16.03	41.59	14.01	59.97	13.75
NWF-CLS	70	45.41	19.86	42.89	18.55	66.96	26.75
NWF-WRC	69	9.26	7.79	8.55	7.34	16.51	12.26
DORF WC	68	30.21	26.30	22.38	24.18	50.56	30.14
RTF	12	17.17	10.02	12.17	12.43	32.33	12.06
				Third Grade			

DORF WC	84	91.73	38.12	89.54	36.92	106.12	36.25
RTF	72	28.18	12.89	32.08	16.65	44.21	15.31
DZC	42	12.46	7.53	15.86	7.59	18.86	7.86
Fourth Grade							
DORF WC	99	110.56	34.06	104.25	28.32	113.71	31.59
RTF	95	28.59	14.52	32.63	20.24	34.47	14.02
DZC	41	17.31	8.50	15.37	6.45	18.49	8.82
Fifth Grade							
DORF WC	100	124.26	38.20	119.75	32.39	127.37	36.91
RTF	97	23.08	12.73	43.02	19.24	45.24	15.16
DZC	61	23.09	8.47	22.93	9.22	23.51	9.34

Note. Middle- and end-of-year complete data. (Time Point A) = middle-of-year benchmark administration with DIBELS 6th Edition directions; (Time Point B) = middle-of-year validation administration with DIBELS Next directions; (Time Point C) = end-of-year benchmark administration with DIBELS 6th Edition directions. DIBELS Next materials used at all time points. FSF = First Sound Fluency. LNF = Letter Naming Fluency; PSF = Phoneme Segmentation Fluency; NWF-CLS = Nonsense Word Fluency Correct Letter Sounds; NWF-WRC = Nonsense Word Fluency Words Read Correctly; DORF WC = DIBELS Oral Reading Fluency Words Corrects; RTF = Retell Fluency; DZC = Daze Adjusted Score. Daze Adjusted Score = Daze number correct - (Daze errors made / 2).

Table 4

DIBELS Measures Quartiles for the Beta 2 Validation Study Sample

	Middle of Year (Time Point A)				Middle of Year Validation (Time Point B)				End of Year (Time Point C)			
Measure	*Q1*	*Q2*	*Q3*	*Q4*	*Q1*	*Q2*	*Q3*	*Q4*	*Q1*	*Q2*	*Q3*	*Q4*
Kindergarten												
FSF	20	31	42	52	19	32	40	49	--	--	--	--
LNF	8	25	33	52	6	22	36	51	23	34.5	45	59
PSF	16	38	48	56	17	33	41	49	39	48	59	71
NWF-CLS	10	16.5	26	40	9	15.5	24	38	20	28	42	71
NWF-WRC	0	0	4	12	0	1	4	11	0	2	9	17
First Grade												
PSF	44	56	63	78	34	44	50	61	53	63	71	77
NWF-CLS	30	45.5	60	81	27	42	53	72	50	65.5	81	121
NWF-WRC	2	9	16	21	2	6	15	21	4	17	23	40
DORF WC	12.5	21	41	96	7	10	25.75	75	26.5	45	74.5	105
RTF	9	16	21	37	3	7.5	20.5	38	24	34.5	40	50

Third Grade												
DORF WC	64	94	122.5	145	62	89.5	119	146	82.5	109.5	132	160
RTF	19	28	38	46	19	29.5	44.5	63	34	43	57.5	70
DZC	7.5	11.25	16	27.5	11	15.5	19	29	15	18	23	33
Fourth Grade												
DORF WC	91	111	135	162	85	104	128	150	94	118	137	164
RTF	18	25	37	55	17	29	48	69	25	33	43	60
DZC	11	16.5	21.5	32	12	14	18	27	12	19	24	32
Fifth Grade												
DORF WC	98.5	123	148	187.5	99.5	118	137	172.5	108	125	148.5	191
RTF	15	21	29	52	28	43	55	76	35	45	56	67
DZC	17	22	27	40	17	21	27	38	17	22	27	40

Note. Middle- and end-of-year complete data. (Time Point A) = middle-of-year benchmark administration with DIBELS 6th Edition directions; (Time Point B) = middle-of-year validation administration with DIBELS Next directions; (Time Point C) = end-of-year benchmark administration with DIBELS 6th Edition directions. DIBELS Next materials used at all time points. Quartile ranks: Q1 = 25th percentile; Q2 = 50th percentile (the median); Q3 = 75th percentile; Q4 = 95th percentile. FSF = First Sound Fluency. LNF = Letter Naming Fluency; PSF = Phoneme Segmentation Fluency; NWF-CLS = Nonsense Word Fluency Correct Letter Sounds; NWF-WRC = Nonsense Word Fluency Words Read Correctly; DORF WC = DIBELS Oral Reading Fluency Words Corrects; RTF = Retell Fluency; DZC = Daze Adjusted Score. Daze Adjusted Score = Daze number correct - (Daze errors made / 2).

Table 5

Percent of Students in each DIBELS Measure Support Category.

	Middle of Year (Time Point A)			Middle of Year Validation (Time Point B)			End of Year (Time Point C)		
Measure	I	S	BM	I	S	BM	I	S	BM
				Kindergarten					
FSF	18.95	15.79	65.26	15.79	16.84	67.37	--	--	--
LNF	31.91	21.28	46.81	35.11	21.28	43.62	34.04	27.66	38.30
PSF	7.29	17.71	75.00	10.42	14.58	75.00	5.21	13.54	81.25
NWF-CLS	10.00	18.89	71.11	10.00	28.89	61.11	11.11	25.56	63.33
				First Grade					
LNF	--	--	--	--	--	--	--	--	--
PSF	0.00	8.96	91.04	1.49	22.39	76.12	0.00	5.97	94.03
NWF-CLS	24.24	34.85	40.91	28.79	36.36	34.85	18.18	36.36	45.45
DORF WC	11.76	32.35	55.88	33.82	27.94	38.24	14.71	23.53	61.76
				Third Grade					
DORF WC	28.57	17.86	53.57	29.76	23.81	46.43	22.62	27.38	50.00
				Fourth Grade					

DORF WC	22.22	16.16	61.62	21.21	29.29	49.49	25.25	24.24	50.51
				Fifth Grade					
DORF WC	19.19	21.21	59.60	20.20	19.19	60.61	20.20	25.25	54.55

Note. Middle- and end-of-year complete data. (Time Point A) = middle-of-year benchmark administration with DIBELS 6th Edition directions; (Time Point B) = middle-of-year validation administration with DIBELS Next directions; (Time Point C) = end-of-year benchmark administration with DIBELS 6th Edition directions. DIBELS Next materials used at all time points. Approximate sample sizes: kindergarten ≈ 94; first grade ≈ 66; third grade = 84; fourth grade ≈ 95; fifth grade ≈ 97. Benchmark Groups: I = Intensive Support; S = Strategic Support; BM = At Benchmark. Benchmark groups based on *DIBELS® 6th Edition Benchmark Goals and Cut-Points for Risk*. Benchmark goals for RTF or Daze had not yet been determined at the time of this study. Percents based on valid scores and do not include students with missing scores. FSF = First Sound Fluency; LNF = Letter Naming Fluency; PSF = Phoneme Segmentation Fluency; NWF-CLS = Nonsense Word Fluency Correct Letter Sounds; DORF WC = DIBELS Oral Reading Fluency Words Correct.

DIBELS Next New Measures: First Sound Fluency and Daze

Research Question #1: What is the alternate-form reliability and the concurrent and predictive validity of the new measures, DIBELS *Next First Sound Fluency and Daze?*

To address Research Question #1, information is presented on FSF and Daze. Descriptive statistics are reported in Table 3. Mean scores are as expected and similar to scores reported in other studies (Cummings, Good, Kaminski, O'Neil, 2010), and Technical Report 11.

To evaluate the reliability, correlations between measures administered at middle-of-year benchmark administration (time point A) and middle-of-year validation administration (time point B) are presented. Concurrent and predictive validity for FSF is presented by correlational data with DIBELS measures LNF, PSF, and NWF with DIBELS Next directions and scoring procedures. Data is presented for Daze as a measure of reading comprehension by correlational data with DORF and RTF.

Alternate-Form Reliability**.** Reliability refers to the relative stability with which a test measures the same skills across minor differences in conditions. The most desirable choice for estimating the reliability of a test is alternate-form reliability with a two-week interval (Nunnally & Bernstein, 1994). Alternate-form reliability indicates the extent to which test results generalize to different item samples, different times, different conditions, and different testers. Students are tested with two different (i.e., alternate) but equivalent forms of the test (preferably at different times, under different conditions, and by different testers) and scores from these two forms are correlated. Alternate-form reliability coefficients may be affected by student learning or practice effects.

Significant differences between alternate-forms are usually interpreted as content sampling error (hence, the requirement for equivalency between forms), but there are additional factors that can affect estimates of test reliability: test length, testing interval, range of student ability in the sample, testing situation, and guessing (Salvia, Ysseldyke, & Bolt, 2007). These factors were addressed and minimized in the following ways.

- *Test Length.* Most DIBELS measures are 1-minute, timed assessments. Generally, students do not complete the form or passage within the allotted time. Ceiling effects are usually not a concern with DIBELS assessments, but floor effects can be observed in the earlier grades.
- *Testing Interval*. Generally, the closer together the administrations, the higher the reliability. Reliability testing was conducted approximately two weeks following middle-of-year benchmark assessment, the preferred amount of time between administrations for alternate-form reliability.
- *Range of Student Ability in the Sample*. When too much or too little variability exists in the sample to provide information on a range of student abilities, the resulting reliability estimates can be inaccurate. The sample for which reliability is estimated was drawn from a fairly low to average-performing population of students.
- *Guessing*. When a student is able to guess, even if the guesses are correct, the responses introduce random error into the score. In an effort to reduce guessing, DIBELS Next measures, with the exception of Daze, employ production-type responses. In addition, students were given opt-out opportunities before validation administrations and were told that they may end the testing session at any time. Students were also encouraged to do their best.
- *Testing Situation*. The student may react to the test (e.g., become frustrated, bored, or lose his/her place). The environment may not be suitable to the student (e.g., the furniture

might be uncomfortable or the room might be cold). These circumstances may introduce an indeterminate amount of error and could lower the reliability of the test. Care was taken to ensure that the student was comfortable within the testing environment, and a rapport was developed between the student and the assessor.

All kindergarten students were administered alternate forms of FSF. All third- through fifth-grade students from a single school where no other alternate-form measures were given were administered alternate forms of Daze. All alternate-form testing was conducted during middle-of-year validation administration (time point B). Alternate-form reliability of a single-form is estimated by the correlation between the score recorded at time point A and the score recorded at time point B. Alternate-form reliability of a three-form aggregate (i.e., for validating need for support in an Outcomes Driven Model or for progress monitoring where a pattern of performance on at least 3 alternate forms is considered) is estimated with the Spearman-Brown Prophecy Formula (Nunnally & Bernstein, 1994). Salvia, Ysseldyke, & Bolt's (2007) standards for reliability were applied; a minimum of .60 is required for administrative purposes and scores that are reported for groups of individuals; a minimum of .80 is required for screening decisions; a minimum of .90 is required for important education decisions concerning an individual student. Results are presented in Table 6.

Table 6

Alternate-Form Reliability Estimates for a Single Assessment and a Three-Form Aggregate for DIBELS Next First Sound Fluency (FSF) and Daze

		Reliability	
Measure	N	Single-Form	Three-Form
		Kindergarten	

FSF	97	.83	.94
Third Grade			
DZC	42	.77	.91
Fourth Grade			
DZC	42	.84	.94
Fifth Grade			
DZC	61	.83	.94

Note. Based on middle of year data only. Reliability coefficients calculated from middle-of-year benchmark administration (time point A) and middle-of-year validation administration (time point B). FSF = First Sound Fluency; DZC = Daze Adjusted Score; DZC = Daze number correct - (Daze errors made / 2). All correlations are significant at the α < .001 level.

The alternate-form reliability of a single-form is .83 for FSF, and .77, .84, and .83 for Daze Adjusted Score in third, fourth, and fifth grades, respectively. These correlations are all significant, *p* < .001. Estimated alternate-form reliability of a three-form aggregate is above .91 for both FSF and Daze. The alternate-form reliability of FSF and fourth- and fifth-grade Daze are above the .80 criterion for screening decisions. Third-grade Daze is arbitrarily close to this cut-point. All estimated three-form reliability coefficients are sufficient for important individual education decisions. These results suggest that DIBELS *First Sound Fluency* and *Daze* are highly reliable measures for use within an Outcomes Driven Model.

Criterion-Related Validity. The concurrent and predictive validity of DIBELS *Next* FSF and data to support Daze as a valid measure of reading comprehension are presented. Hopkins (2002) standards for validity are applied; very small correlational relationships are less than .09, small correlational relationships range from .10 - .29, moderate correlational relationships range from .30 - .49, moderate-strong correlational relationships range from .50 - .69, and strong correlational relationships are above .70. The concurrent and predictive validity of FSF is presented through correlations with DIBELS measures PSF and NWF in Table 7. Data is presented for Daze as a measure of reading comprehension through correlations with DORF and RTF in Table 8.

Table 7

Concurrent and Predictive Validity for Middle of Year DIBELS Next First Sound Fluency

	Middle of Year Measures (Concurrent, Time Point A)	End of Year Measures (Predictive, Time Point C)		
Measure	PSF	PSF	NWF-CLS	NWF-WRC
FSF (B)	.74*** (96)	.53*** (96)	.34***(96)	.25* (.92)

Note. Correlations are based on subjects with pair-wise complete data, which are reported in parentheses. (Time Point A) = middle-of-year benchmark administration with DIBELS 6th Edition directions; (Time Point B) = middle-of-year validation administration with DIBELS Next directions; (Time Point C) = end-of-year benchmark administration with DIBELS 6th Edition directions. DIBELS Next materials used at all time points. PSF = Phoneme Segmentation Fluency; NWF = Nonsense Word Fluency; CLS = Correct Letter Sounds; WRC = Words Read Completely and Correctly. Significant codes: '***' p < .001; '*' p < .05.

Table 8

Correlational Relationships of Middle-of-Year DIBELS Next Daze

	Daze Adjusted Score by Time of Year	
Middle of Year Measure (Time Point A)	Middle of Year (Concurrent, Time Point A)	End of Year (Predictive, Time Point C)
	Third Grade	
DORF Words Correct	.72 (124)	.81 (127)
Retell Fluency	.39 (115)	.40 (116)
Daze Adjusted Score	--	.73 (124)

Fourth Grade		
DORF Words Correct	.75 (139)	.74 (140)
Retell Fluency	.31 (135)	.31 (135)
Daze Adjusted Score	--	.75 (139)
Fifth Grade		
DORF Words Correct	.75 (161)	.75 (161)
Retell Fluency	.31 (157)	.27 (157)
Daze Adjusted Score	--	.79 (161)

Note. Correlations are based on subjects with pair-wise complete data. The number with pair-wise complete data is reported in parentheses. DORF and Retell Fluency were administered with DIBELS 6th Edition directions and scoring procedures. Daze administered with DIBELS Next directions and scoring procedures. DIBELS Next materials used at all time points. DORF = DIBELS Oral Reading Fluency. Daze Adjusted Score = Daze number correct - (Daze errors made / 2). All correlations are significant at the α < .001 level.

Concurrent validity coefficients for FSF range from .39 (NWF-WRC) to .74 (PSF) indicating moderate to strong correlational relationships. Across third, fourth, and fifth grades, correlations between DORF Words Correct and Daze adjusted score are strong; all correlations are above the .70 criterion. The correlational relationships of middle-of-year Daze to end-of-year Daze is also strong; all correlations are above the .70 criterion for all grades. The correlations of Retell Fluency (RTF) with Daze range from .38 to .40 in third grade, indicating moderate correlational relationships, and range from .26 to .31 in fourth and fifth grades, indicating small to moderate correlational relationships.

These results indicate that FSF is a valid measure of early phonemic awareness, and that Daze is a valid measure of reading comprehension.

The Changes to Directions

Research Question #2: What are the middle-of-year intercorrelational relationships and predictive validity to end-of-year DIBELS outcomes of DIBELS measures Letter Naming Fluency, Phoneme-Segmentation Fluency, and Nonsense Word Fluency with the modified directions and scoring procedures in development for DIBELS Next?

Research Question #3: What is the effect, if any, of changes to directions and scoring procedures for DIBELS *Letter Naming Fluency, Phoneme-Segmentation Fluency, and Nonsense Word Fluency?*

To address Research Questions #2 and #3, information is presented on the effect of changes to the directions for DIBELS Next measures LNF, PSF, and NWF. Information about DORF, which was administered with the DIBELS Next directions, is also presented as a criterion measure.

The revisions to the directions appeared in the scoring booklets and the *DIBELS Beta 2 Administration and Scoring Guide* (ASG). Tables 9, 10, and 11 present both versions of the directions that appear in the examiner booklets (DIBELS 6th Edition directions are from DIBELS 6th Edition booklets and DIBELS Next directions are from DIBELS Beta 2 Validation Study booklets) for DIBELS Next LNF, PSF, and NWF. As mentioned earlier in this report on page 14, directions labeled as "DIBELS Next directions" were directions and scoring procedures in development for DIBELS Next.

Table 9

Directions for DIBELS Letter Naming Fluency (LNF)

Category	DIBELS 6th Edition Directions	DIBELS Next Directions
Introduction	***Here are some letters*** (point to the student probe). ***Tell me the names of as many letters as you can.***	***I'm going to show you some letters. I want you to point to each letter and say its name.*** (Put the page of letters in front of the student.)
Begin testing prompts	***When I say, "Begin," start here*** (point to first letter), ***and go across the page*** (point). ***Point to each letter and tell me the name of that letter. If you come to a letter you don't know I'll tell it to you. Put your finger on the first letter. Ready, begin.***	***Start here.*** (Point). ***Go this way*** (Sweep your fingers across the first two rows of letters) ***and say each letter name. Put your finger under the first letter*** (point). ***Ready, begin.***
Timing	1 minute. Start your stopwatch after telling the student to begin.	1 minute. Start your stopwatch after telling the student to begin.
Wait	Allow 3 seconds, provide the correct letter. Point to the next letter and say, ***"What letter?"***	Allow 3 seconds, score the letter incorrect, provide the correct letter and if necessary, point to the next letter.
Discontinue	No letters named correctly in the first row.	No letters named correctly in the first row.

Reminders	***"Remember to tell me the letter name, not the sound it makes."*** (Allowed 1 time.)	Student does not go left to right: ***Go this way.*** (Allowed 1 time.) Student skips 4 consecutive letters: ***Try to say <u>each</u>*** **letter name. (Allowed 1 time.)** Student says letter-sounds: ***Say the letter <u>name</u>, not its sound.*** (Allowed 1 time.) If student stops (and it's not a hesitation on a specific item): ***"Keep going."*** If student loses his/her place, point.

Note. Directions are from DIBELS 6th Edition and DIBELS Beta 2 Validation Study scoring booklets <u>*only*</u>. Scoring booklets had the administration scripts along with a note before the introduction to each measure that said, "Make sure you have reviewed the long form of the directions in the *DIBELS Administration and Scoring Guide* and have them available." Directions are formatted as they appear on the scoring guides.

Table 10

Directions for DIBELS Phoneme Segmentation Fluency (PSF)

Category	DIBELS 6th Edition Directions	DIBELS Next Directions
Introduction	***I am going to say a word. After I say it, you tell me all the sounds in the word. So, if I say, "sam," you would say /s/ /a/ /m/. Let's try one*** (one-second pause). ***Tell me the sounds in "mop."***	***We are going to say the sounds in words. Listen to me say all the sounds in the word fan. /f/ /a/ /n/. Listen to another word:*** (pause) ***jump. /j/ /u/ /m/ /p/. Your turn. Say all the sounds in "soap".***
Begin testing	***OK. Here is your first word.***	***I'm going to say more words. I will say the word and***

prompt		***you say all the sounds in the word. Ready? _____.***
Timing	Give the student the first word and start your stopwatch.	1 minute. Start your stopwatch after saying the first test item.
Wait	Allow 3 seconds, then give the student the next word.	Allow 3 seconds, then give the student the next word.
Discontinue	No segments produced correctly in the first 5 words.	No segments produced correctly in the first 5 words.
Reminders	ASG	If student spells the word: ***Say the sounds in the word.***(Allowed 1 time.) If student repeats the word: ***Remember to say all the sounds in the word.*** (Allowed 1 time.)

Note. Directions are from DIBELS 6th Edition and DIBELS Beta 2 Validation Study scoring booklets *only*. ASG = Directions were provided in the *DIBELS 6th Edition Administration and Scoring Guide* only, and do not appear in the scoring booklet. Scoring booklets had the administration scripts along with a note before the introduction to each measure that said, "Make sure you have reviewed the long form of the directions in the *DIBELS Administration and Scoring Guide* and have them available." Directions are formatted as they appear on the scoring guides.

Table 11

Directions for DIBELS Nonsense Word Fluency (NWF)

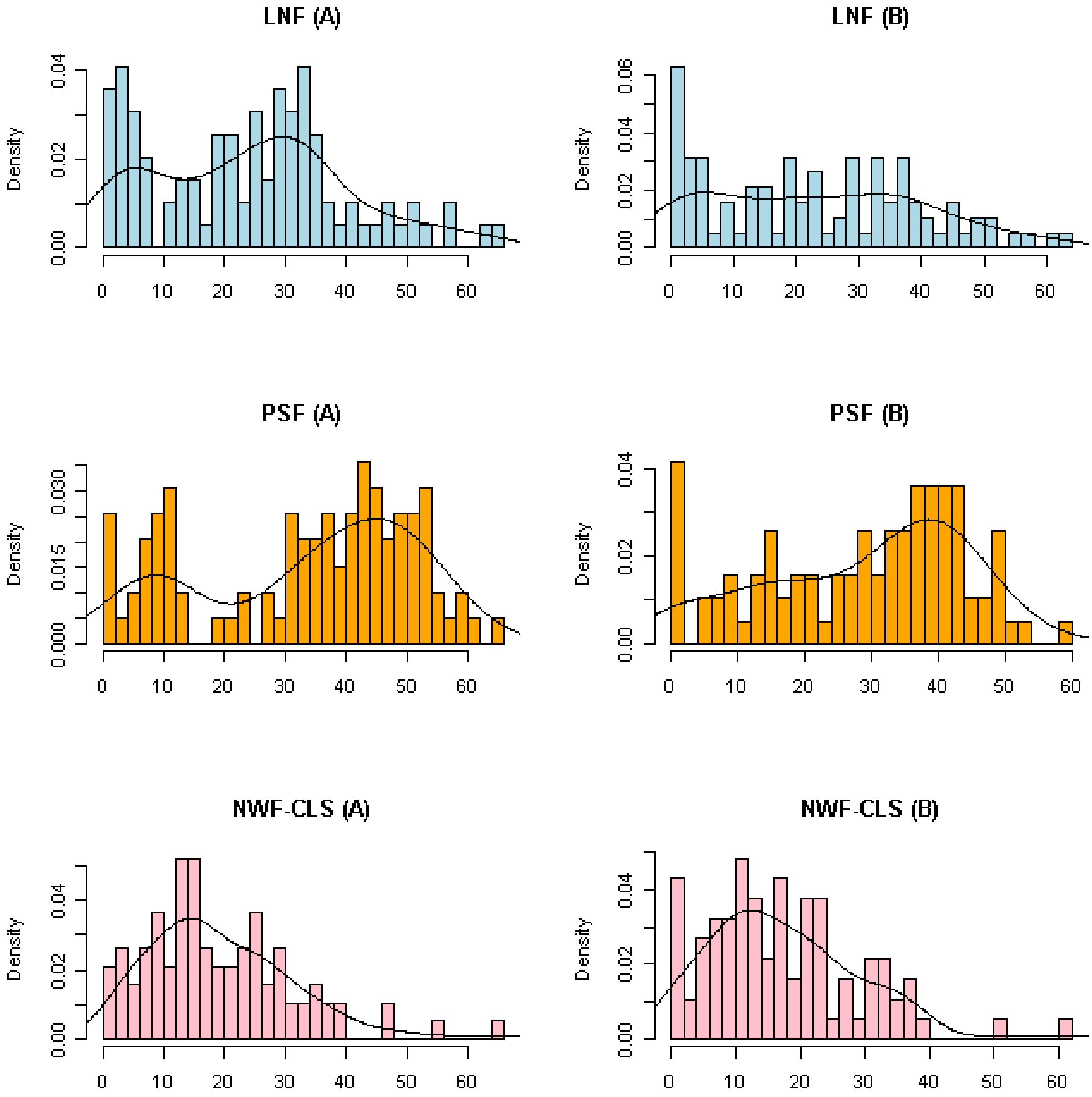
Category	DIBELS 6th Edition Directions	DIBELS Next Directions

Introduction	***Look at this word*** (point to the first word on the practice probe). ***It’s a make-believe word. Watch me read the word: /s/ /i/ /m/, “sim”*** (point to each letter then run your finger fast beneath the whole word). I ***can say the sounds of the letters, /s/ /i/ /m/*** (point to each letter), ***or I can read the whole word, “sim”*** (run your finger fast beneath the whole word). ***Your turn to read a make-believe word. Read this word the best you can*** (point to the word “lut”). ***Make sure you say any sounds you know.***	***We are going to read some make-believe words.*** (Place the sample copy in front of the student). ***Listen. This word is “sog.”*** (Run your finger under the word as you say it.) ***The sounds are /s/ /o/ /g/*** (point to each letter). ***Your turn. Read this make-believe word*** (point to the word “mip”). ***If you can’t read the whole word, tell me any sounds you know.***
Begin testing prompt	Place the student copy of the probe in front of the child. ***Here are some more make-believe words*** (point to the student probe). ***Start here*** (point to the first word) a***nd go across the page*** (point across the page). ***When I say, “Begin,” read the words the best you can. Point to each letter and tell me the sound or read the whole word. Read the words the best you can. Put your finger on the first word. Ready, begin.*** Start your stopwatch.	***I'd like you to read more make-believe words. Do your best reading. If you can’t read the whole word, tell me any sounds you know.*** (Place the student probe in front of the student.) ***Put your finger under the first word. Ready, begin.***
Timing	1 minute. Start your stopwatch after telling the student to begin.	1 minute. Start your stopwatch after telling the student to begin.
Wait	Sound-by-Sound: Allow 3 seconds, then provide the correct letter sound. Point to the next word and say, ***"What word?"*** Word by Word: Allow 3 seconds, then provide the correct word. Point to the next word and say, ***"What word?"***	Responding sound-by-sound, mixing sounds and words, or sounding out and recoding: Allow 3 seconds, then provide the correct letter sound. Responding with whole words: allow 3 seconds, then provide the correct word.
Discontinue	No sounds correct in the first row.	No sounds correct in the first row.

Reminders	ASG		Student does not go left to right: ***Go this way.*** (Allowed 1 time.) Student says letter names: ***Say the sounds, not the letter names.*** (Allowed 1 time.) Student reads the word first, then says the letter-sounds: ***Just read the word.*** Student says all of the letter sounds correct in the first row, but does not make any attempt to blend or recode: ***Try to read the words as whole words.*** If student stops (and it's not a hesitation on a specific item): ***"Keep going."*** If student loses his/her place, point.

Note. Directions are from DIBELS 6th Edition and DIBELS Beta 2 Validation Study scoring booklets only. ASG = directions were provided in the *DIBELS 6th Edition Administration and Scoring Guide* only, and do not appear in the scoring booklet. Scoring booklets had the administration scripts along with a note before the introduction to each measure that said, "Make sure you have reviewed the long form of the directions in the *DIBELS Administration and Scoring Guide* and have them available." Directions are formatted as they appear on the scoring guides.

Distribution Histograms for Kindergarten DIBELS Next Measures Letter Naming Fluency, Phoneme Segmentation Fluency, and Nonsense Word Fluency Correct Letter Sounds



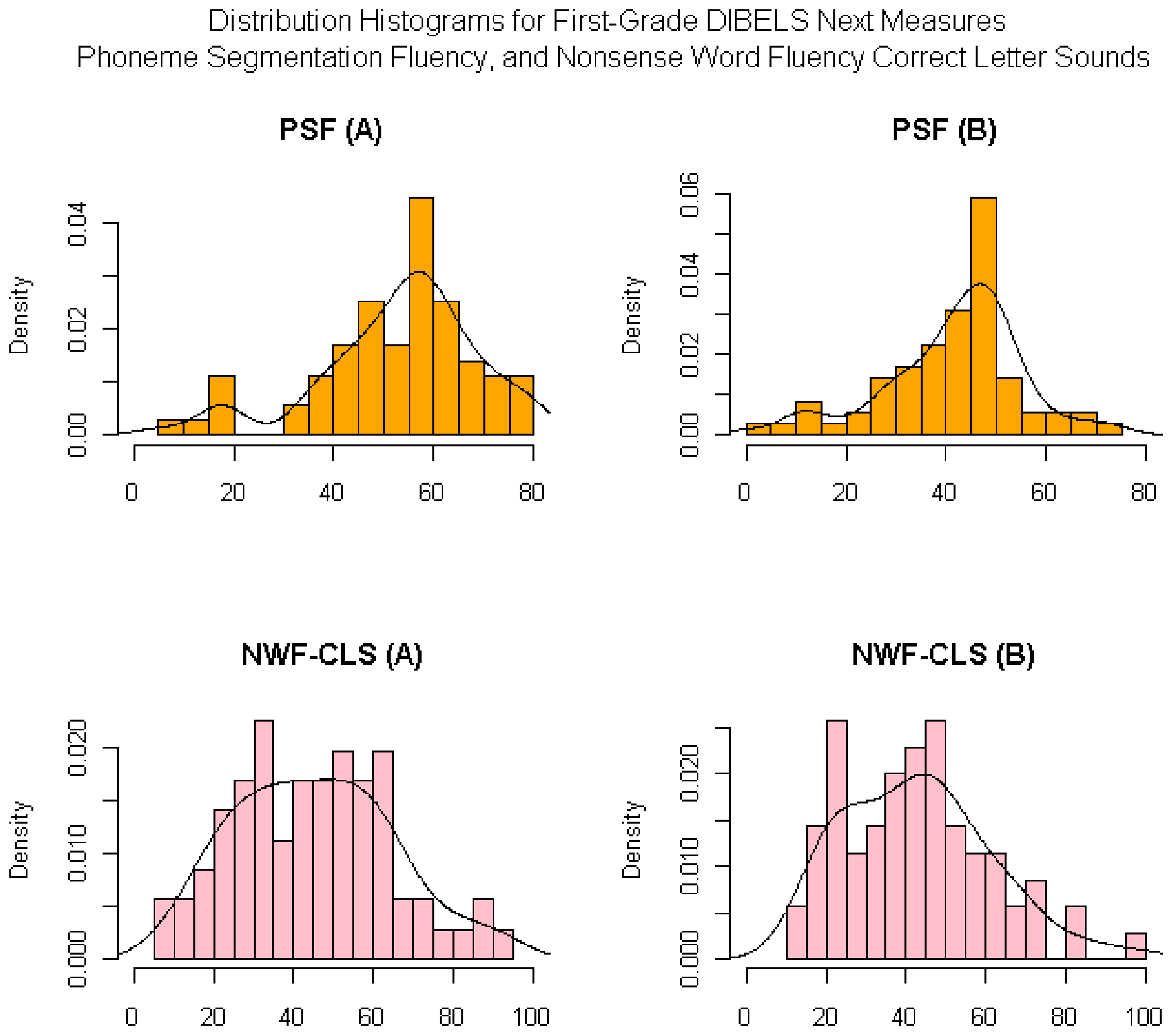


Figure 1. Histograms and fitted locally weighted scatterplot smoothing (lowess) lines for scores from middle of year kindergarten and first-grade DIBELS measures. A = middle of year benchmark administration, B = middle of year validation administration. LNF = Letter Naming Fluency; PSF = Phoneme Segmentation Fluency; NWF-CLS = Nonsense Word Fluency Correct Letter Sounds.

Descriptive Statistics. Descriptive statistics for LNF, PSF, NWF-CLS, DORF, and RTF for kindergarten and first grade are reported in Tables 3 and 4. Distribution histograms LNF, PSF, and NWF-CLS with both sets of directions are presented in Figure 1.

Floor effects were observed on all kindergarten measures (from Figure 1). Ceiling effects are observed in first-grade PSF at middle of year benchmark administration (time point A) and end-of-year benchmark administration (time point C, histogram not shown). The histograms of scores for both kindergarten and first-grade PSF suggests a bimodal distribution at time point A, in which a separate group of students scored very low while the remaining students scored in the expected range. Figure 1 suggests a practice effect for PSF, evident by this group of low-achieving students scoring higher at time point B, and thus the distribution of kindergarten PSF appearing unimodal with a zero-spike and skewed to the left. There are similar results for kindergarten LNF.

All students generally performed the same at time point A than time point B (Table 3). Mean scores from middle of year benchmark assessment (time point A) fall within similar ranges to middle-of-year validation administration (time point B), except for first-grade PSF and DORF WC, in which there are sharp decreases. Compared to mean scores from time point A, at time point B, kindergarten students scored 2 points fewer, on average, across LNF, PSF, and NWF-CLS, and first-grade students scored 3 points fewer on NWF-CLS and 12 points fewer on PSF. Mean scores for PSF are higher than expected; over 75% of kindergarten and first-grade students scored above the benchmark for PSF (Table 5). Variability in scores increases across all grades between benchmark administrations, with the exception of PSF.

Tests about Mean, Variance, and Correlation**.** To assess the effects of changes to directions on scores for LNF, PSF, and NWF, tests about the mean, variance, and correlation were performed. Tests that are robust against departures from normality were chosen, because of the floor effects, ceiling effects, and excess of zero scores in the data. Where appropriate, *p*-values and effect sizes are reported.

For tests about means, Wilcoxon Signed-Rank tests were performed to test for equality of means between two measures. Cohen's *d* is the effect size reported for tests about means and is calculated as the difference in mean scores between two measures divided by their pooled standard deviation. Hopkins recommends guidelines for interpreting *d*: trivial ($0 < d < .2$), small ($.2 < d < .6$), moderate ($.6 < d < 1.2$), moderate-strong ($1.2 < d < 2$), strong ($2 < d < 4$), and nearly perfect ($4 < d$).

For tests about variance, Levene's tests were performed to test for homogeneity of variance. The percentage of variance shared by two variables is the effect size reported for tests about variance. The percentage of variance explained is calculated by squaring the correlation coefficient, *r*. Hopkins recommends guidelines for interpreting the amount of shared variance: small (0-.30), moderate (.30-.50), moderate-strong (.50-.70), strong (.70-.90), nearly perfect (greater than .90).

For tests about correlation, Fisher's Z-transformation was performed to test for equality of correlations. Cohen's *q* is the effect size reported for tests about correlation. Cohen's *q* is the difference in z-transformed correlation coefficients between two sets of two variables, and assesses the degree to which we support the hypothesis of equal correlation. The absolute value of this metric, $|q|$, is reported. Cohen provides guidelines for interpreting *q*; small ($|q| = .10$), medium ($|q| = .30$), and large ($|q| = .50$).

A summary of statistical test results and effect sizes for differences about means (Cohen's *d*) and variances (percent of variance explained) are reported in Table 12 and about correlation (Cohen's *q*)in Table 13.

Table 12

Statistical Test Results and Effect Sizes for Tests about the Mean and Variance between DIBELS 6th Edition and DIBELS Next Measures LNF, PSF, and NWF-CLS with Different Directions

		Tests about the Mean			Tests about the Variance		
Measure by Grade	N	*diff*	*Effect Size*	*p*	*F*	*Effect Size*	*p*
Kindergarten							
LNF	94	0.43	0.03	.76	1.08	.74	.30
PSF	96	4.92	0.30	.02	2.15	.77	.14
NWF-CLS	90	1.77	0.15	.29	0.00	.61	.95
First Grade							
PSF	71	11.39	0.76	.00	0.70	.35	.41
NWF-CLS	70	2.53	0.13	.45	0.59	.61	.45

Note. Based on data from time points A and B. (Time Point A) = middle-of-year benchmark administration with DIBELS 6th Edition directions; (Time Point B) = middle-of-year validation administration with DIBELS Next directions. DIBELS Next materials used at all time points. Wilcoxon Signed-Rank tests were performed to test for equality in means, and Levene's tests were performed to test for homogeneity of variance. Cohen's *d* (smaller is better) is reported for the effect size for differences in means and the percentage of shared variance (larger is better) is reported for the effect size for differences in variance.

Table 13

Statistical Test Results and Effect Sizes for Tests about Correlation between DIBELS 6th Edition and DIBELS Next Measures LNF, PSF, and NWF-CLS with Different Directions

	Tests about Correlation			
Measure by Grade	Time Point Combination	*r*	*Effect Size*	*p*
Kindergarten				
LNF	A with C	.85*** (94)		
	B with C	.85*** (94)	.14	.20
PSF	A with C	.72*** (96)		
	B with C	.69*** (96)	.03	.77
NWF-CLS	A with C	.72*** (90)		
	B with C	.76*** (90)	.10	.33
First Grade				
PSF	A with C	.59*** (71)		
	B with C	.39*** (71)	.27	.03
NWF-CLS	A with C	.63*** (70)		
	B with C	.67*** (70)	.09	.45

Note. Correlations are based on subjects with pair-wise complete data. The number with pair-wise complete data is reported in parentheses. (Time Point A) = middle-of-year benchmark administration with DIBELS 6th Edition directions; (Time Point B) = middle-of-year validation administration with DIBELS Next directions; (Time Point C) = end-of-year benchmark administration with DIBELS 6th Edition directions. DIBELS Next materials used at all time points. Fisher's Z-transformation was used to test for equality of correlation. Cohen's |*q*| (smaller is better) is reported for the effect size. Significant code: '***' *p* < .001.

In Table 12, the results from paired Wilcoxon Signed-Rank tests for differences in means (see Table 3 for descriptive statistics) between middle-of-year measures with different directions indicate significant differences for kindergarten PSF (*d* = 0.30, *p* = .02) and first-grade PSF (*d* = 0.76, *p* < .001). Levene's tests for differences in variance between middle-of-year measures with different directions yielded non-significant results for all measures. In Table 13, tests about correlation between measures with different directions yielded non-significant results for all measures except first-grade PSF (q = -0.27, p = .03).

Tests about the Intercorrelations between DIBELS Measures. The relationship between middle-of-year LNF, PSF, and NWF for kindergarten and first grade is represented by intercorrelations with measures from the same time point; e.g., if X = correlation between scores from LNF and PSF (middle of year benchmark administration), and Y = correlation between scores from VLNF and VPSF (middle of year validation administration), then we test for differences between X and Y. The intercorrelations between middle-of-year administrations (time points A and B) is reported in Tables 14 and 15.

Tests about Predictive Validity. The predictive validity is the relation between an earlier construct and the same construct or a later construct at later points in time. The predictive validity of middle-of-year LNF, PSF, and NWF for kindergarten and first grade is represented by intercorrelations with end-of-year measures. All end-of-year DIBELS measures were administered using DIBELS 6th Edition directions. Scores from both middle-of-year benchmark administration (time point A) and middle-of-year validation administration (time point B) are correlated with scores from end-of-year benchmark administration (time point C), and are reported in Tables 16 and 17.

Table 14

Comparison of Intercorrelations of Kindergarten Middle-of-Year DIBELS Measures with Different Directions.

	Criterion Measure		
Measure by Administration	PSF	NWF-CLS	NWF-WRC
Letter Naming Fluency (LNF)			

6th Ed. Directions (A)	.59*** (98)	.75*** (96)	.55*** (96)
Next Directions (B)	.64*** (95)	.66*** (91)	.52*** (89)
Phoneme Segmentation Fluency (PSF)			
6th Ed. Directions (A)	--	.52*** (96)	.40*** (96)
Next Directions (B)	--	.51*** (93)	.47*** (91)
Nonsense Word Fluency Correct Letter Sounds (NWF-CLS)			
6th Ed. Directions (A)		--	.72*** (96)
Next Directions (B)		--	.75*** (91)

Note. Based on data from time points A and B. (Time Point A) = middle-of-year benchmark administration with DIBELS 6th Edition directions; (Time Point B) = middle-of-year validation administration with DIBELS Next directions. DIBELS Next materials used at all time points. Correlations are based on subjects with pair-wise complete data. The number with pair-wise complete data is reported in parentheses. Significant code: '***' *p* < .001.

Table 15

Comparison of the Intercorrelations of First-Grade Middle-of-Year DIBELS Measures with Different Directions.

	Criterion Measure		
Measure by Administration	NWF-CLS	NWF-WRC	DORF WC
Phoneme Segmentation Fluency (PSF)			
6th Ed. Directions (A)	.40* (71)	.30*(70)	.13† (71)
Next Directions (B)	.32**(70)	.40***(70)	.30*(70)
Nonsense Word Fluency Correct Letter Sounds (NWF-CLS)			
6th Ed. Directions (A)	--	.73*** (70)	.74*** (71)
Next Directions (B)	--	.67***(70)	.67***(70)
Nonsense Word Fluency Words Read Completely and Correctly (NWF-WRC)			
6th Ed. Directions (A)		--	.66*** (70)

Next Directions (B)	--	.88***(70)

Note. Based on data from time points A and B. (Time Point A) = middle-of-year benchmark administration with DIBELS 6th Edition directions; (Time Point B) = middle-of-year validation administration with DIBELS Next directions. DIBELS Next materials used at all time points. Correlations are based on subjects with pair-wise complete data. The number with pair-wise complete data is reported in parentheses. Significant codes: '***' *p* < .001; '**' *p* < .01; '*' *p* < .05; '†' *p* > .05, i.e. not significant.

Table 16

Comparison of the Predictive Validity of Kindergarten Middle-of-Year DIBELS Measures with Different Directions.

Measure by Assessment Period	End of Year DIBELS 6th Edition Directions (Time Point C)			
	LNF	PSF	NWF-CLS	NWF-WRC
Letter Naming Fluency (LNF)				
6th Ed. Directions (A)	.85***(97)	.47***(97)	.66***(97)	.56***(93)
Next Directions (B)	.81***(94)	.42***(94)	.61***(94)	.48***(90)
Phoneme Segmentation Fluency (PSF)				
6th Ed. Directions (A)	--	.72***(97)	.45***(97)	.37***(93)
Next Directions (B)	--	.69***(96)	.45***(96)	.38***(92)
Nonsense Word Fluency Correct Letter Sounds (NWF-CLS)				
6th Ed. Directions (A)	--	--	.72***(95)	.62***(91)
Next Directions (B)	--	--	.76***(92)	.63***(89)
Nonsense Word Fluency Words Read Completely and Correctly (NWF-WRC)				
6th Ed. Directions (A)	--	--	.66***(95)	.74***(91)
Next Directions (B)	--	--	.73***(86)	.74***(88)

Note. (Time Point A) = middle-of-year benchmark administration with DIBELS 6th Edition directions; (Time Point B) = middle-of-year validation administration with DIBELS Next directions; (Time Point C) = end-of-year benchmark administration with DIBELS 6th Edition directions. DIBELS Next materials used at all time points. Correlations are based on subjects with pair-wise complete data. The number with pair-wise complete data is reported in parentheses. Significant code: '***' *p* < .001.

Table 17

Comparison of the Predictive Validity of First-Grade Middle-of-Year DIBELS Measures with Different Directions.

Measure by assessment period	End-of-Year DIBELS 6th Edition Directions (Time Point C)		
	NWF-CLS	NWF-WRC	DORF WC
Phoneme Segmentation Fluency (PSF)			
6th Ed. Directions (A)	.32**(71)	.27*(71)	.25*(71)
Next Directions (B)	.27*(71)	.24*(71)	.34**(71)
Nonsense Word Fluency Correct Letter Sounds (NWF-CLS)			
6th Ed. Directions (A)	.63***(71)	.58***(71)	.68***(71)
Next Directions (B)	.67***(70)	.62***(70)	.67***(70)
Nonsense Word Fluency Words Read Completely and Correctly (NWF-WRC)			
6th Ed. Directions (A)	.60***(70)	.67***(70)	.61***(70)
Next Directions (B)	.56***(70)	.70***(70)	.60***(70)

Note. (Time Point A) = middle-of-year benchmark administration with DIBELS 6th Edition directions; (Time Point B) = middle-of-year validation administration with DIBELS Next directions; (Time Point C) = end-of-year benchmark administration with DIBELS 6th Edition directions. DIBELS Next materials used at all time points. Correlations are based on subjects with pair-wise complete data. The number with pair-wise complete data is reported in parentheses. DORF = DIBELS Oral Reading Fluency Words Correct. Significant codes: '***' *p* < .001; '**' *p* < .01; '*' *p* < .05.

Tests for differences in the intercorrelational relationships between middle of year DIBELS LNF, PSF, and NWF (Tables 14 and 15) returned non-significant results for all comparisons. Non-significant effect sizes (and *p*-values) ranged from -.02 (*p* = .86) to .24 (*p* = .05); the smallest effect size (*q* = -.02) is reported for kindergarten measures middle-of-year PSF with middle-of-year NWF-CLS, and the largest effect size (*q* = .24) is reported for first-grade measures middle-of-year NWF-WRC with middle-of-year DORF.

Test for differences in predictive validity (Tables 16 and 17) returned non-significant results for all comparisons. Non-significant effect sizes (and *p*-values) ranged from .01 (*p* = .96) to -.14 (*p* = .21); the smallest effect size (*q* = .01) was reported for kindergarten measures middle-of-year PSF with end-of-year NWF-CLS, and the largest effect size (*q* = -.14) was reported for kindergarten measures middle-of-year NWF-WRC with end-of-year NWF-CLS.

The validation administration of PSF in the middle of the year (time point B) using the DIBELS Next directions provided scores that were significantly lower than the corresponding scores from the district-administered PSF (time point A) for both kindergarten and first grade (Table 12, *d* = 0.76, *p* < .001). The correlation with the end of year district-administered PSF with 6th edition directions also was lower (Table 13, *q* = -.27, *p* = .03). However, the variances of scores were not significantly different for the validation administration with Next directions and district-administered 6th directions (Table 12, percent of variance explained = 35%, *p* = .71), and the predictive validity coefficients with later constructs were not significantly different (Table 17, NWF-CLS, *q* = .06, *p* = .65; NWF-WRC, *q* = .04, *p* = .80; and DORF WC, *q* = -.10, *p* = .42).

These results concerning first-grade PSF were not expected, and thus we looked for evidence in the data that would explain why these results tell such a different story than previous studies. We outline three possibilities, starting with the most plausible, and ending with the least plausible.

1. The observed high achievement on middle-of-year benchmark PSF may have caused statistical tests to wrongly detect a significant difference when there is none; i.e., high

achievement may have inflated the Type I error rate. This explanation is the most plausible; more than 70% of students were above benchmark at all three time points for DIBELS Next PSF.

2. The significant differences are due to time sampling error; i.e., two much time elapsed between administrations, and students increased their knowledge about the skills assessed in DIBLES Next PSF, thus their scores changed. This explanation is possible but unlikely; Salvia and Ysseldyke recommend a two-week time frame for alternate-form administrations, and this short amount of time is unlikely to produce significant changes in scores.
3. The significant differences could be attributable to content sampling error (i.e., the alternate-form of PSF at time point B was not equivalent in form to the PSF administered at time points A and C); however, both forms were constructed from the same word pool and followed the same format, so this explanation is also improbable.

Point 1 explores differences in PSF through student achievement, and points 2 and 3 explore differences in PSF through content and administration. The most plausible explanations for the differences in PSF are attributable to student achievement, and therefore judgment on whether the different directions for DIBELS Next PSF impacted the measure are inconclusive.

Nonsense Word Fluency Scores

Research Question #4: How do DIBELS *Nonsense Word Fluency scores Words Read Completely and Correctly (WRC), words Sounded Out and Recoded (SOR), and Whole Words Read (WWR) compare to each other and contribute to the NWF measure?*

To answer Research Question #4, information was collected on NWF strategy type. A recent study by Harn, Stoolmiller, and Chard (2008) exploring the relationship between reading skill and reading strategy suggested that students who attempt to read nonsense words on DIBELS NWF as whole words performed better on DORF than students who utilized other reading strategies, such as sounding out words or partially blending sounds together. During middle-of-year validation administration, DMG data collectors recorded two extra scores for NWF to further explore this finding. In addition to correct letter sounds (CLS) and words read completely and correctly (WRC), a score indicating the number of nonsense words sounded out and recoded (SOR) and a score indicating the number of words read as whole words (WWR) were also recorded. WRC is equivalent to the sum of SOR and WWR. For this study, we examined whether WWR is more highly related to other DIBELS Next outcomes than SOR or WRC.

***Descriptive Statistics, Intercorrelations, and Validity*.** Table 18 presents descriptive statistics for NWF score types from middle-of-year validation administration (time point B). Intercorrelations between NWF score types are presented in Table 19. The correlational relationships of WRC, SOR, and WWR with NWF-CLS and DORF is presented in Table 20. As described on page 27 of this report, Hopkins (2002) standards of validity are applied.

Table 18

Descriptive Statistics and Quartiles for Kindergarten and First-Grade Nonsense Word Fluency (NWF) Scores Administered at Time Point B

Score	N	*Mean*	*SD*	*Q1*	*Q2*	*Q3*	*Q4*
			Kindergarten				
WRC	91	2.79	3.83	0	1	4	11
SOR	91	1.85	2.61	0	0	3	8
WWR	91	0.95	2.89	0	0	0	6

			First Grade				
WRC	70	8.43	7.36	2	6	15	21
SOR	70	1.79	3.64	0	0	2	10
WWR	70	6.64	7.24	1	3	13	20

Note. Based on data from time point B (middle-of-year validation data with DIBELS Next directions and scoring procedures). Quartile ranks: Q1 = 25th, Q2, = 50th, Q3 = 75th, and Q4 = 95th. WRC = Words Read Correctly; SOR = Number of Words Sounded Out and Re-coded; WWR = Whole Words Read.

Table 19

Intercorrelations for Middle-of-Year Nonsense Word Fluency (NWF) Scores Administered at Time Point B

NWF Score	WRC	SOR	WWR
	Kindergarten		
CLS	.75***	.39***	.64***
WRC	--	.66***	.73***
SOR		--	-.03†
WWR			--
	First Grade		
CLS	.67***	.03†	.67***
WRC	--	.28*	.88***
SOR		--	-.22†
WWR			--

Note. Based on data from time point B (middle-of-year validation data with DIBELS Next directions and scoring procedures). Correlations are based on subjects with pair-wise complete data. Sample sizes: Kindergarten = 91, first grade = 70. WRC = Words Read Correctly; SOR = Words Sounded Out and Re-coded; WWR = Whole Words Read; CLS = Correct Letter Sounds. Significant codes: '***' p < .001; '**' p < .01; '*' p < .05; '†' p > .05, i.e. not significant.

Table 20

Intercorrelations for Nonsense Word Fluency Scores

NWF Score at Time Point B	Middle of Year Validation (Time Point B)		End of Year (Time Point C)	
	NWF-CLS	DORF WC	NWF-CLS	DORF WC
		Kindergarten		
WRC	.75***(91)	--	.74***(90)	--
SOR	.39***(91)	--	.27*(90)	--
WWR	.64***(91)	--	.74***(90)	--
		First Grade		
WRC	.67***(70)	.50***(68)	.56***(70)	.60***(70)
SOR	.03†(70)	.04†(68)	.07†(70)	.17†(70)
WWR	.67***(70)	.49***(68)	.53***(70)	.53***(70)

Note. Based on data from time points B and C. (Time Point B) = middle-of-year validation administration with DIBELS Next directions; (Time Point C) = end-of-year benchmark administration with DIBELS 6th Edition directions. DIBELS Next materials used at all time points. Correlations are based on subjects with pair-wise complete data. The number with pair-wise complete data is reported in parentheses. WRC = Words Read Correctly; SOR = Words Sounded Out and Re-coded; WWR = Whole Words Read; CLS = Correct Letter Sounds; DORF WC = DIBELS Oral Reading Fluency Words Correct. Significant codes: '***' $p < .001$; '**' $p < .01$; '*' $p < .05$; '†' $p > .05$, i.e. not significant.

The number of kindergarten students that successfully read nonsense words as whole words is low; the average WWR score is approximately 1 (Table 18). By first grade, the average number of nonsense words read as whole words increased to 7. The average number of nonsense words sounded out and recoded (SOR) stayed approximately the same between kindergarten and first grade (1.85 in kindergarten, and 1.79 in first grade).

The score types WRC and WWR share a large portion of variability (r = .73 in kindergarten, and r = .88 in first grade from Table 19). The correlation between SOR and WRC shrinks from .66 in kindergarten to .28 in first grade (see Table 19). Since WRC is a linear combination of WWR and SOR, then the small correlation with SOR and the large correlation with WWR suggests that the correlational relationships associated with WRC are affected by WWR and not SOR.

WWR and CLS share a large portion of variability as well (*r* = .64 and .67 in first grade and *r* = .74 and .53 in first grade at middle and end of year, respectively, from Table 20). The correlations between SOR and CLS are smaller first-grade correlations are not statistically significant (*r* = .39 and .27 in kindergarten and *r* = .03 and .07 in first grade for middle and end of year, respectively, from Table 20). The results suggest that WWR is a more important skill and pattern of performance than SOR (and thereby WRC).

***Regression Analysis*.** A regression analysis was performed to distinguish between groups of students who read words as whole words from those students who did not. Students were separated into two groups based on whether or not the student scored higher than the median score for each SOR (median = 0 in kindergarten, and 2 in first grade) and WWR (median = 0 in kindergarten, and 3 in first grade). Residual analysis suggested that the assumptions behind the regressions were not violated.

Figure 2 box plots illustrate the differences in WWR from kindergarten to first grade relative to SOR and WRC. Figure 3 scatter plots with a locally weighted scatterplot (lowess) smoothing line illustrate the linear relationship between WWR and CLS. Table 21 presents estimates and *p*-values from a regression predicting two different end-of-year reading outcomes. Middle-of-year NWF scores with DIBELS Next directions were used to predict performance on end-of-year CLS and DORF WC with DIBELS 6th Edition directions.

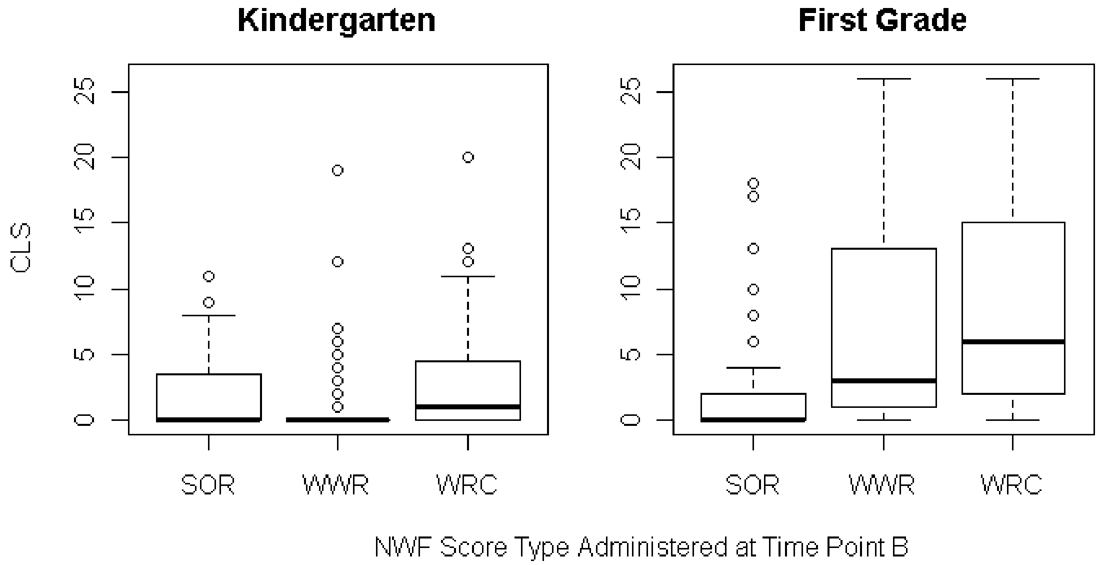


Figure 2. Box plots of *Nonsense Word Fluency* (NWF) score type by grade. V= Measures administered with DIBELS Next directions and scoring procedures; CLS = NWF correct letter sounds, SOR = NWF sounded out and recoded; WWR = NWF whole words read; WRC = NWF words read completely and correctly.

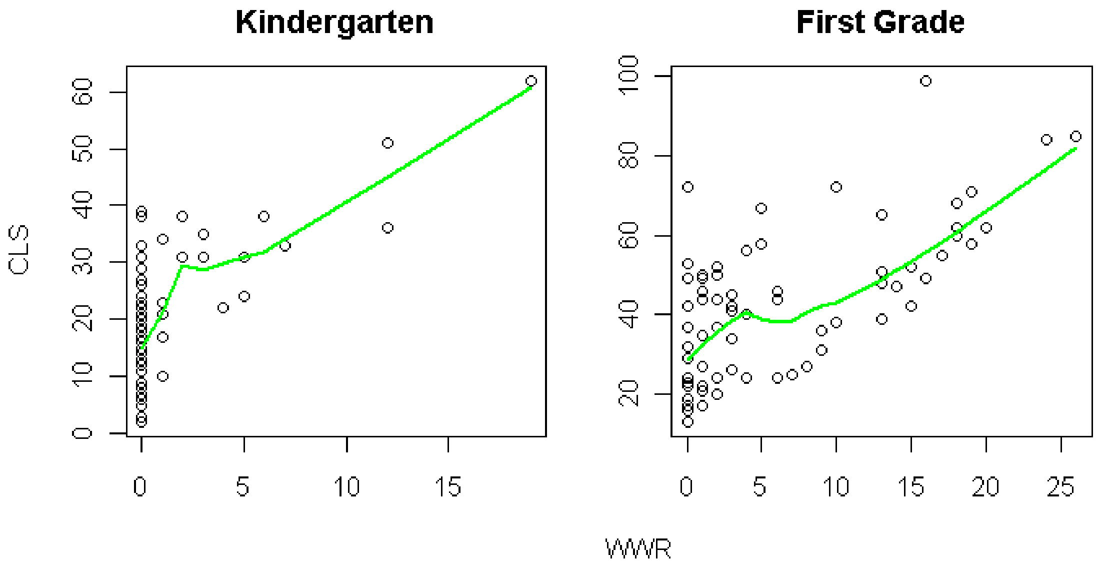


Figure 3. Scatter plots of Nonsense Word Fluency Whole Words Read (NWF-WWR) with Nonsense Word Fluency Correct Letter Sounds (NWF-CLS) by grade with lowess smoothing line. Score ranges are different between grades. V= Measures administered with DIBELS Next directions and scoring procedures.

The average WWR score changed significantly between grades ($p < .001$), but the average SOR score did not ($p = .91$), as illustrated in the box plots from Figure 2. Approximately 80% of kindergarten students did not read any words whole. In first grade, this proportion shrunk to 23%. In Figure 3, the lowess smoothing line illustrates a significant increasing relationship between reading words as whole words and a higher score on NWF-CLS. Between grades, and on different ranges of scores, the pattern of increase is the same. These results suggest that WWR increases substantially from kindergarten to first grade and SOR does not.

Table 21

Predicting End-of-Year Reading Outcomes from Middle-of-Year (Time Point B) Nonsense Word Fluency Scores and Scoring Groups.

		Model 1		Model 2	
		Response = NWF-CLS (Time Point C)		Response = DORF WC (Time Point C)	
Parameter / Measure from Time Point B	N	*Estimate*	*p*	*Estimate*	*p*
		Kindergarten			
Intercept		11.25	.00	--	--
CLS		1.10	.00	--	--
SOR Score ≥ 1	42	1.24	.68	--	--
WWR Score ≥ 1	18	7.97	.06	--	--
		First Grade			
Intercept		24.39	.00	4.43	.52
CLS		0.78	.00	0.84	.00
SOR Score ≥ 3	30	5.18	.27	4.68	.37
WWR Score ≥ 4	34	14.21	.01	17.85	.00

Note. Based on data from time points B and C. (Time Point B) = middle-of-year validation administration with DIBELS Next directions; (Time Point C) = end-of-year benchmark administration with DIBELS 6th Edition directions. DIBELS Next materials used at all time points. Sample sizes: Kindergarten = 93, first grade = 71; First Grade = 71. Model 1 R-Square = .59 for kindergarten and .51 for first grade. Model 2 R-Square = .51 for first grade. SOR and WWR Groups separate students based on if they scored above the median on SOR and WWR. In first grade, the median scores are 0 and 0, respectively. In first grade, the median scores are 2 and 3, respectively. DORF WC = DIBELS Oral Reading Fluency Words Correct; CLS = Correct Letter Sounds; SOR = Words Sounded Out and Recoded; WWR = Whole Words Read.

There is convincing evidence that first-grade students who score higher than the median on WWR (median = 3) obtain higher scores on end-of-year NWF-CLS and DORF (*p* = .01, and *p* < .005, respectively). There is suggestive but inconclusive evidence that kindergarten students see similar benefits with a higher score on WWR (*p* = .06). On average, a first-grade student who reads at least 4 nonsense words as whole words will score 14 points higher on end-of-year CLS and 18 points higher on end-of-year DORF. These findings are consistent with the findings from Harn et al., (2008), and suggest that reading nonsense words as whole words is a strong indicator of future reading skill.

Exploratory Factor Analysis / Principle Components Analysis. Exploratory factor analysis (EFA) coefficients are estimates of the amount of common variance present in the data. Principle components analysis (PCA) coefficients are estimates of the maximum variability shared among the variables. Both analyses' components form a model that relates the components to a latent immeasurable skill that influences responses on the variables (Child, 1990). There are assumptions behind both analyses. EFA assumes there exists a causal structure that accounts for the shared variability among the measures. PCA is simply a variable reduction procedure that attempts to maximize the variability in scores explained by the components with the fewest components possible. The model coefficients (i.e., final communality estimates) represent the proportion of the variance associated with that variable that is both error-free and shared with the other variables in the model. A large coefficient (greater than .70) indicates a strong relationship with all other coefficients in the model, and therefore a strong relationship to an underlying causal structure. Thus, the measures that identify strongly with this underlying causal structure will maximize model variance and individual component variance. While it is not always appropriate to evaluate data with both procedures, EFA provides a useful conceptual model for interpreting PCA (Ramsey & Schafer, 2002).

Reading proficiency, as a construct, is not directly measureable. There are many contributing factors to a student's reading skill. To examine the contribution of NWF scores

WRC, SOR and WWR in kindergarten and first grade, a PCA and EFA were performed with each score in a separate model with measures FSF, LNF, PSF, and NWF-CLS. Each model was examined separately, and factors were retained based on commonly used guidelines: positive eigenvalues, percentage of variance explained, a scree test, the size of the residuals, and interpretability (Nunnally & Bernstein, 1994, Kim & Mueller, 1978, Jolliffe, 2002). Only DIBELS measures with DIBELS Next directions were included in the analysis.

Tables 22 and 23 present the results from the PCA and EFA for kindergarten and first grade, respectively.

Table 22

Principal and Factor Component Variance Estimates for Kindergarten Nonsense Word Fluency (NWF) Scores

	Model 1		Model 2		Model 3	
Measure / Score	PCA	EFA	PCA	EFA	PCA	EFA
First Sound Fluency	.87	.60	.62	.59	.81	.60
Letter Naming Fluency	.68	.57	.67	.61	.69	.58
Phoneme Segmentation Fluency	.85	.73	.76	.74	.85	.72
NWF Correct Letter Sounds	.87	.73	.58	.53	.84	.73
NWF Words Read Correctly	.82	.63	--	--	--	--
NWF Sounded Out and Recoded	--	--	.38	.25	--	--
NWF Whole Words Read	--	--	--	--	.85	.49
Model Variance	.81	.65	.60	.55	.81	.62

Note. N = 91. All measures were administered during middle-of-year validation administration with DIBELS Next materials and directions (time point B). "Model Variance" is the average of the component variance estimates, and represents the percentage of total variance present among all measures. PCA = Principal Components Analysis; EFA = Exploratory Factor

Analysis.

Table 23

Principal and Factor Component Variance Estimates for First Grade Nonsense Word Fluency (NWF) Scores

	Model 1		Model 2		Model 3	
Measure / Score	PCA	EFA	PCA	EFA	PCA	EFA
Phoneme Segmentation Fluency	.94	.37	.58	.58	.96	.27
DORF Words Correct	.80	.64	.78	.51	.76	.81
NWF Correct Letter Sounds	.81	.74	.83	.84	.82	.56
NWF Words Read Correctly	.75	.67	--	--	--	--
NWF Sounded Out and Recoded	--	--	.70	.07	--	--
NWF Whole Words Read	--	--	--	--	.70	.61
Model Variance	.82	.60	.72	.50	.81	.56

Note. N = 70. All measures were administered during middle-of-year validation administration with DIBELS Next materials and directions (time point B). "Model Variance" is the average of the component variance estimates, and represents the percentage of total variance present among all measures. DORF = DIBELS Oral Reading Fluency. PCA = Principal Components Analysis; EFA = Exploratory Factor Analysis.

The estimate of model variance is the average of the component variance estimates, and represents the percentage of total variance present among all measures. Thus, the model within each grade that captures the majority of the variability associated with the underlying causal structure is the model that best fits.

In both kindergarten and first grade, Model 1 has the largest communality estimates, as expected. The score for WRC is the sum of SOR and WWR, thus capturing more information and more variability. The estimates from Model 3 are similar to Model 1, which suggests that the majority of the variability in WRC is captured in WWR. Model 2 explains the least variance, and the large discrepancy in the estimates for first-grade SOR in Model 2 raises some legitimate concerns about whether or not SOR belongs in the model. These results suggest that WWR is the source component of WRC that shares variability with other DIBELS measures.

Discussion

What is the alternate-form reliability and the concurrent and predictive validity of the new DIBELS NEXT measures *First Sound Fluency* and *Daze*? The criterion for screening decisions is .80. The alternate-form reliability of FSF is above this criterion, as is fourth and fifth-grade Daze. Third-grade Daze is arbitrarily close to this cut-point. All estimated reliability coefficients for three-form aggregates are sufficient for important individual education decisions. For example, in validating need for support decisions, students might be retested 3 times with alternate forms to increase confidence in the decision to provide support. In progress monitoring decisions, the pattern of performance on 3 or more alternate forms administered over time might be used to evaluate progress. These results suggest that DIBELS *First Sound Fluency* and *Daze* are highly reliable measures for use within an Outcomes Driven Model.

Using Hopkins (2002) standards for validity, the concurrent validity of FSF is moderate to strong with other measures of early phonemic awareness. Correlations between DORF and Daze adjusted score are strong. These results support the validity of FSF as a measure of early phonemic awareness, and support the validity of Daze as a measure of reading comprehension.

What are the intercorrelational relationships and the predictive validity of DIBELS measures *Letter Naming Fluency*, *Phoneme-Segmentation Fluency*, and *Nonsense Word Fluency* with the modified directions and scoring procedures in development for DIBELS Next? The intercorrelational relationships with other middle of year measures and the predictive validity of LNF and NWF is moderate to strong, and the intercorrelational relationships and predictive validity of PSF is small to moderate-strong with other measures of early literacy skills and phonemic awareness The results indicate that LNF, NWF, and kindergarten PSF are valid measures of early literacy skills and phonemic awareness. There appears to be a drop-off in the validity coefficients for PSF in first grade, likely due to student mastery (92% of first-grade students are at or above benchmark by end of year, see Table 5).

What is the effect, if any, of changes to directions and scoring procedures for DIBELS *Letter Naming Fluency*, *Phoneme-Segmentation Fluency*, and *Nonsense Word Fluency*? The answer to this research question is limited to what the data can prove. The data will only prove that the directions did *not* alter the measures. Tests that reveal significant differences could be due to the change in directions, but also could be due to practice effects, content sampling error, and error associated with changes over time. There is no evidence to suggest that the change in directions altered the outcome of the kindergarten measures LNF, PSF, or NWF-CLS and first-grade NWF-CLS. Although the mean scores from kindergarten PSF and NWF-CLS were significantly lower with the changes to directions, the variance, correlation with end-of-year measures, and the middle-of-year intercorrelational relationships and the predictive validity remained the same. These results suggest that the change in directions did not introduce a substantial new source of variability into the data for these measures, and did not significantly alter the correlational relationships with other DIBELS measures. The significant differences in first-grade PSF between benchmark administration (time points A and C) and validation administration (time point B) with different directions are assumed to be a natural consequence of ceiling effects and student mastery of phonemic segmentation.

How do DIBELS *Nonsense Word Fluency* scores *Words Read Completely and Correctly* (WRC), words *Sounded Out and Recoded* (SOR), and *Whole Words Read* (WWR) compare to each other and contribute to the NWF measure? The score representing the number of words read as whole words (NWF-WWR) is the component of NWF-WRC that relates strongly to other DIBELS measures. Students who have a higher score on NWF-WWR score higher on NWF-CLS and DORF. NWF-WWR possesses score stability and validity, and is a good indicator of future reading skill.

Limitations. When evaluating the percent of students at or above benchmark (Table 5), the average-achieving cut-point is approximately 60%; a larger percentage represents a 'high-achieving' group of students, and below 50% represents a 'low-achieving' group of students. This sample was low-achieving on kindergarten LNF, first-grade NWF-CLS, and third- and fourth-grade DORF WC; average-achieving on kindergarten FSF, NWF-CLS, and sixth-grade DORF WC; and high-achieving on both kindergarten and first-grade PSF. Data were collected in a single school district from a small town in the Pacific Northwest, and thus inference should be limited to this population and/or populations that this school district adequately represents. In addition, knowledge of instructional context is limited, although we do know that the participating school district has a dedicated program of early literacy instruction with resources for those with special needs (vision, hearing, attendance, etc.) and environmental supports.

Implications. The new measures FSF and Daze appear to be working well in this sample. Previous work has documented the reliability and validity of these measures which is illustrated in this study as well. NWF-WWR appears to function well with NWF-CLS and DORF. Correlations among all measures are strong, indicating the measures are accurately evaluating specific reading components and predicting future outcomes on similar measures. Finally, for more information about the benchmark goals for these measures, or further information about their validity, please refer to Technical Reports 7 and 11, the DIBELS 6th Edition Technical Adequacy Information (2008), and the DIBELS Next Technical Manual (2011).

References

Bloom, Richburg-Hayes, and Black (2007). Using Covariates to improve precision for studies that randomize schools to evaluate educational interventions. *Educational Evaluation and Policy Analysis*, *29*, 30-59.

Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research,* 1, 245-276.

Child, D. (1990). *The essentials of factor analysis* (2nd ed.). London: Cassel Educational Limited.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Cummings, K. D., Kaminski, R. A., Good, R. H., & O'Neil, M. (2010). Assessing phonemic awareness in preschool and kindergarten: development and initial validation of First Sound Fluency. *Assessment for Effective Intervention*.

Dale, E., & O'Rourke, J. (1976). *The Living Word Vocabulary: The Words We Know: A National Vocabulary Inventory.* Elgin, IL: Field Enterprises Educational Corporation.

Darlington, R. B., (2004). Factor Analysis. Retrieved Nov, 2010, from http://comp9.psych.cornell.edu/Darlington/factor.htm

Factor Analysis Using SAS PROC FACTOR. UCLA: Academic Technology Services, Statistical Consulting Group. from http://www.ats.ucla.edu/stat/sas/library/factor_ut.htm (accessed December 2, 2010).

Good III, R. H., & Kaminski, R. A. (Eds.). (2002). Dynamic Indicators of Basic Early Literacy Skills (6th ed.). Eugene, OR: Institute for the Development of Educational Achievement. Available: http://dibels.uoregon.edu/.

Good III, R. H., Kaminski, R. A., Cummings, K., Dufour-Martel, C., Petersen, K., Powell-Smith, K., Stollar, S., & Wallin, J. (2010) DIBELS Next Assessment Manual. From https://dibels.org/next/ (accessed Dec 10, 2010).

Harn, B.A., Stoolmiller, M., & Chard, D.J. (2008). Measuring the dimensions of alphabetic principle on the reading development of first graders: the role of automaticity and unitization. *Journal of Learning Disabilities*, 41(2), 143-157.

Hatcher, L. (1994). *A Step-by-Step Approach to Using the SAS System for Factor Analysis and Structural Equation Modeling*. Cary, NC: SAS Institute, Inc.

Hedges, L., Hedberg, E. (2007). Intraclass correlation values for planning group-randomized trials in educational research. *Educational Evaluation and Policy Analysis*, 29, 60-87.

Hopkins, W. G. (2002). A scale of magnitudes for the effect statistics. In *A review of statistics*. Retrieved July 30, 2010 from http://www.sportsci.org/resource/stats/effectmag.html.

Hotelling, H. (1933) Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24:417-441,498-520.

Jolliffe, I.T., (2002). *Principal Component Analysis* (2nd ed.). New York: Springer-Verlag New York, Inc.

Kaminski, R.A., Baker, S.K., Chard, D., Clarke, B., Smith, S. (2006). *Final report: Reliability, Validity, and Sensitivity of Houghton Mifflin Early Growth Indicators* (Technical Report). Eugene, OR: Dynamic Measurement Group and Pacific Institutes for Research

Kaminski, R. A., Good, R. H. (1996). Toward a technology for assessing basic early literacy skills. *School Psychology Review*, 25, 215-227.

Kim, J. O., & Mueller, C. W. (1978). *Factor analysis: Statistical methods and practical issues.* Newbury Park, CA: Sage.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). New York: McGraw-Hill.

Powell-Smith, K. A., Good, R. H., & Atkins, T. (2010). *DIBELS Next Oral Reading Fluency Readability Study* (Technical Report No. 7). Eugene, OR: Dynamic Measurement Group.

Ramsey, F. L., & Schafer D. W. (2002). *The Statistical Sleuth: a course in methods of data analysis* (2nd ed.). Pacific Grove, CA: Duxbury.

Salvia, J., Ysseldyke, J., & Bolt, S. (2007). *Assessment in special and inclusive education* (10th ed). Boston: Houghton Mifflin Company.

SAS Institute Inc., SAS 9.1.3 Help and Documentation, Cary, NC: SAS Institute Inc., 2002-2004.

SAS Institute Inc., SAS OnlineDoc 9.1.3, Cary, NC: SAS Institute Inc., 2002-2005.

Slavin, R. (2008). What works? Issues in synthesizing educational program evaluations. *Educational Researcher*, 37, 5-14.

StatSoft, Inc. (2010). "Reliability and Item Level Analysis." In: *Electronic Statistics Textbook.* Tulsa, OK: StatSoft. WEB: http://www.statsoft.com/textbook/ (accessed January 13, 2010).

Tabachnick, B. G., Fidell, L. S. (2007). *Using Multivariate Statistics* (5th ed). Boston: Allyn and Bacon.

U.S. Dept. of Education, National Center for Education Statistics, CCD (2007-08). Demographic Information for a single school district. Retrieved October 20, 2010 from http://nces.ed.gov/.