Technical Adequacy of DIBELS: Results of the Early Childhood Research Institute on measuring growth and development

Roland Good
Ruth Kaminski
Mark Shinn
John Bratten
Michelle Shinn
Debby Laimon
Sylvia Smith
Natalie Flindt

University of Oregon

Citation

Good, R.H., Kaminski, R.A., Shinn, M., Bratten, J., Shinn, M., Laimon, D., Smith, S., & Flindt, N. (2004). *Technical Adequacy of DIBELS: Results of the Early Childhood Research Institute on measuring growth and development* (Technical Report, No. 7). Eugene, OR: University of Oregon.

Author Note

This research was supported in part by the Early Childhood Research Institute on Measuring Growth and Development (U.S. Department of Education H024360010). 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 Roland H. Good III, School Psychology Program, College of Education, University of Oregon. Eugene, OR 97405.

Abstract

The Early Childhood Research Institute on Measuring Growth and Development (ECRI-MGD) examined reliability and validity of the Dynamic Indicators of Basic Early Literacy Skills (DIBELS) in a four year, longitudinal research study across 10 cohorts of children in kindergarten through third grade. The purpose of this study was to examine whether the DIBELS measures are reliable and valid indicators of children's early literacy skills and whether the DIBELS are effective tools for monitoring the individual progress of students. Four DIBELS measures were examined in the study: (a) Initial Sound Fluency (ISF), (b) Phoneme Segmentation Fluency (PSF), (c) Letter Naming Fluency (LNF), and (d) Nonsense Word Fluency (NWF).

Each student was assessed monthly with all DIBELS measures throughout the study. Therefore, multiple reliability and validity coefficients are reported. The ECRI-MGD examined in particular the median reliability, concurrent validity, and long-term predictive validity at specific target times for each measure. For kindergartners and first graders, all DIBELS measures displayed adequate reliability. When 3 or 4 (ISF) probes are aggregated together, all DIBELS measures have estimated reliability in the .90s. The median concurrent validity of single DIBELS probes with the Woodcock-Johnson Broad Reading Cluster were .36 for ISF, .56 for PSF, .51 for NWF, and .75 for LNF. The DIBELS measures were also found to predict both oral reading fluency (ISF median r = .38, PSF median r = .62, NWF median r = .69) and Woodcock Johnson Total Reading Cluster score (ISF median r = .33, PSF median r = .63, NWF median r = .66) more than a year later. Implications of these findings are discussed.

Technical Adequacy of DIBELS: Results of the Early Childhood Research Institute on measuring growth and development

A report from the National Assessment of Educational Progress (NAEP) states that "approximately 40% of students across the nation cannot read at a basic level." (U.S. Department of Education, 2003, p. #1). This is a grave social concern because low reading skills are related to delinquency, school dropout rates and unemployment (McGill-Franzen, 1987). In order for children to experience success in academic, social and economic outcomes, it is imperative that children make adequate progress in learning to read at an early age (Good, Simmons, & Smith, 1998). Children on low developmental reading trajectories experience significant difficulty "catching up" to their peers (Good et al., 1998). For example, a study by Juel (1988, p. 437) found the "probability that a child would remain a poor reader at the end of fourth grade if the child was a poor reader at the end of first grade was .88." In contrast, Juel also found the "probability that a child would remain an average reader in fourth grade if the child had average reading ability in first grade was .87" (p. 440).

These disturbing findings have alerted the nation to the need for more attention on the early reading achievement of all children. The No Child Left Behind (NCLB) Act of 2001 was developed to ensure that every child can read by the end of third grade. To accomplish this goal, the Reading First Initiative was created to provide funds to assist states with professional development, selection of effective instructional materials, and administration of reading assessments (U.S. Department of Education, 2002). The recipients of a Reading First grant are required to administer scientifically-based screening and diagnostic assessments to determine which kindergarten through third grade students are at-risk for reading difficulties. Furthermore, school districts must

monitor the progress of students and show evidence that their reading program is effective by assessing important reading outcomes.

Due to these requirements from the NCLB and Reading First Initiative, schools can no longer depend exclusively on the end-of-third-grade, state-wide assessments to determine which students are at-risk. Instead, educators must use valid, reliable, and formative assessments to identify early at-risk students and to monitor a child's progress on critical early literacy skills.

The National Reading Panel report (2000) conducted a meta-analysis of the research literature pertaining to critical early literacy skills. Their analysis indicated there are five foundational reading skills strongly and causally related to reading outcomes. These include phonemic awareness, alphabetic principle, fluency, reading comprehension and vocabulary development. It is essential that schools assess *all* students on these early literacy skills to identify those who are at-risk and intervene early to prevent future reading failure.

One way to ensure that all students are on track for being successful readers is to make assessment decisions within an Outcomes-Driven Model. The Outcomes-Driven Model accomplishes steps to outcomes through a set of five educational decisions: (a) identifying need for support, (b) validating need for support, (c) planning and implementing support, (d) evaluating and modifying support, and (e) reviewing outcomes (Good, Gruba, & Kaminski, 2002). This step-by step process ensures that all children are given the support necessary to become successful readers in a proactive, preventive, early intervention framework.

For assessments to be effective in informing educational decisions that will prevent reading failure, they must be reliable and valid measures of risk as well as brief

and repeatable for on-going progress monitoring and outcome evaluation. These requirements are set forth in the NCLB and Reading First Initiative, and they are necessary to ensure the reading success of all students.

The Dynamic Indicators of Basic Early Literacy Skills (DIBELS) is one assessment tool that has been developed for use within an Outcomes-Driven Model to identify children at-risk for reading difficulties and monitor students' progress throughout their academic instruction (Good, Gruba, & Kaminski, 2001). The DIBELS assessment includes measures that assess the five foundational early literacy skills of children in kindergarten through third grade. The measures include Initial Sound Fluency (ISF), Phoneme Segmentation Fluency (PSF), Nonsense Word Fluency (NWF) and Letter Naming Fluency (LNF), and, subsequent to this study, DIBELS Oral Reading Fluency (DORF).

The present research study was conducted by The Early Childhood Research Institute on Measuring Growth and Development (ECRI-MGD). The ECRI-MGD, established in 1996, was funded by the Office of Special Education and Rehabilitation Services (OSERS), U.S. Department of Education to research and develop a comprehensive measurement system to assess the individual needs of children birth to eight years of age and their families (McConnell et al., 1998). One reason for developing a measurement system for young children was the increased attention to accountability for academic *outcomes* of children rather than educational *processes* (McConnell et al., 1998; Priest et al., 2001).

A second reason was the need for an assessment system that would measure the growth and development of early childhood skills across a broad age range that allows an evaluation of intervention effectiveness for an individual child in addition to groups of

children. Assessing the progress of individual children is necessary to monitor children's progress and to evaluate an intervention's effectiveness (Priest et al., 2001).

This study examines the technical adequacy of the DIBELS, which is a measurement system developed to assess the early literacy skills of children kindergarten through third grade. The ECRI-MGD institute examined longitudinal data across four years to answer the following questions: Are the DIBELS measures reliable and valid indicators of children's early literacy skills? Are the DIBELS effective tools for monitoring the individual progress of students on those early reading skills?

Method

Participants

Participants were from kindergarten, first, second, and third grade classrooms in two elementary schools in separate school districts. Data were collected from the first elementary school beginning in the 1997-98 academic year, and data were collected at the second elementary school beginning in the 1998-99 academic year. Both schools remained in the study until its conclusion at the end of the 2000-01 academic year. Most participants were followed for multiple consecutive years; thus, ECRI-MGD could base findings regarding literacy and language growth and development on across-grade, within-subject longitudinal data.

Participant selection. The sample of students was obtained in a multi-step process. Once ECRI-MGD was granted access to a school by the school's principal, individual teachers volunteered their classrooms as sources of participants. Once a classroom entered the study, informed consent documents were distributed to each student's parents, and students were enrolled in the study once their parents' consent was provided. When new students entered a participating classroom, their parents were given

an opportunity to enroll them in the study. As students' families moved into and out of the area, the study's sample size continuously fluctuated (see Table 1).

Parents provided informed consent for their children to participate in the study over multiple years. When possible, ECRI-MGD followed participants from one grade level to the next. At the beginning of each academic year, students who moved from a non-participating classroom to a participating classroom were eligible to enter the existing cohort of participants. Students who went from a participating classroom to one whose teacher refused to participate were lost to the study. Each year a new cohort of participants was formed in each school as new students entered the participating kindergarten classrooms.

Composition of cohorts. ECRI-MGD followed cohorts of students through as many grade levels as possible over the course of four academic years. Each participant was assigned membership in a specific cohort in accordance with their classroom placement for the year they entered the study. Cohorts are titled with two-digit numbers. The first digit in a cohort title represents school number ("1" for School 1 and "2" for School 2). The second digit distinguishes a given cohort from others based on grade level and date of the entire cohort's entry into the study. For example, at School 1 in 1997-98, ECRI began data collection with kindergartners (Cohort 11) and first-graders (Cohort 12) in participating classrooms. Both of these cohorts remained in the study through third grade; thus, there were four waves of data collection for Cohort 11 and three waves for Cohort 12. If, in 1999-2000 a student entered a participating second grade classroom and was granted her parent's consent to participate in the study, she would have become a member of Cohort 11. If she were a year older and entered a participating third grade

classroom she would become a member of Cohort 12. Table 1 summarizes the years of data collection and length of involvement for each cohort.

Longitudinal data. For three of these cohorts, ECRI-MGD has longitudinal data that spans grades kindergarten through second grade, and for one cohort there are data from kindergarten through third grade. Due to student mobility, the number of participants in each cohort for whom ECRI-MGD ultimately has complete longitudinal data across all study years is small relative to the yearly average numbers shown in Table 1. By the end of the 2000-2001 academic year, however, between the two schools, there were roughly 40 participants in each grade level from whom there are complete or nearly complete (missing data from two or fewer data collection points) longitudinal data across all study years. For most purposes, however, longitudinal data spanning more than two years are not necessary, and the sample sizes for most analyses reflected in this report are generally larger than 40.

Demographic information. Both schools involved in the current study are located in Lane County, Oregon, and are near the city of Eugene, the second most populous urban area in the state. In the 2000-2001 academic year, School One had a total of about 490 students. It is located in the eighth largest city in Oregon (population 52,864). School Two, in contrast, has a total of about 580 students during 2000-2001 and is located in the 71st largest city in the state (population 4,721). In terms of location, both schools are characterized by the National Center for Education Statistics (NCES) as urban fringe of mid-size city. See Table 2 for detailed information about ethnicity and socio-economic status of students in these schools.

Table 1
Length of Participation, Years of Data Collection, and Median Cohort Size for Each Cohort Involved in DIBELS Data Collection for the Early Childhood Research Institute

	Academic year			
Cohort	1997-1998	1998-1999	1999-2000	2000-2001
		School 1		
Cohort 11	Y1 - Kindergarten ($n \cong 81$)	Y2 - First grade ($n \cong 92$)	Y3 - Second grade ($n \cong 88$)	Y4 - Third grade ($n \cong 76$)
Cohort 12	Y1 - First grade ($n \cong 75$)	Y2 - Second grade ($n \cong 78$)	Third grade ($n \cong 83$)	
Cohort 13		Y1 - Kindergarten ($n \cong 75$)	Y2 - First grade ($n \cong 85$)	Y3 - Second grade ($n \cong 82$)
Cohort 14			Y1 - Kindergarten ($n \cong 66$)	Y2 - First grade ($n \cong 79$)
Cohort 15				Y1 - Kindergarten ($n \cong 81$)
		School 2		
Cohort 21		Y1 - Kindergarten ($n \cong 41$)	Y2 - First grade ($n \cong 72$)	Y3 - Second grade ($n \cong 84$)
Cohort 22		Y1 - First grade ($n \cong 65$)	Y2 - Second grade ($n \cong 87$)	Y3 - Third grade ($n \cong 109$)
Cohort 23		Y1 - Second grade ($n \cong 61$)	Y2 - Third grade ($n \cong 108$)	
Cohort 24			Y1 - Kindergarten ($n \cong 94$)	Y2 - First grade ($n \cong 98$)
Cohort 25				Y1 - Kindergarten ($n \cong 64$)

Note. Annual median sample sizes are given for each cohort at each wave of data collection. Sample sizes fluctuated continuously due to constant enrollment of new students and attrition of others. Each year of data collection included multiple assessments across the academic year.

Table 2
Ethnicity and Socio-economic Status of Students in Schools One and Two Compared with that of the General Population of the United States, Oregon, and Lane County.

	Comparable regions			Study sites	
Demographic	United States	Oregon	Lane County	School 1	School 2
Ethnicity					
Native American	0.9	1.3	1.1	1.0	0.3
Asian/Pacific Islander	3.7	3.2	2.2	2.6	0.9
Hispanic	5.5	4.2	1.9	7.3	3.6
Black, non-Hispanic	12.3	1.6	0.8	1.2	0.9
White, non-Hispanic	75.1	86.6	90.6	87.8	94.3
Two or more races	2.4	3.1	3.3	n/a	n/a
Proportion of low-income students[a]	--	--	--	40.7	41.9

Note. School level data were taken from the National Center for Education Statistics and the Oregon Department of Education. Country, state, and county level data were taken from 2000 Census data.
[a]As measured by proportion of students eligible for the USDA Free and Reduced Lunch Program.

Measures

Dynamic Indicators of Basic Early Literacy Skills (DIBELS) and Curriculum-Based Measurement of Oral Reading Fluency (CBM ORF) comprised the DIBELS assessment materials at the time of this study. The DIBELS benchmark and progress-monitoring materials are available for free download to registered users at *http://dibels.uoregon.edu*. Users are requested to register to document usage and to provide a way to alert users to modifications, revisions, and additions to the DIBELS materials. Once users have downloaded and printed a copy of the assessment materials, that copy is used as a photocopy master to create sufficient assessment materials for the

school or district. Also available at the DIBELS web site is the DIBELS Data System, a data entry and reporting service currently available on a fee for service basis. DIBELS Data System users can enter scores using a web browser and obtain the class and school reports illustrated in Good, Gruba, and Kaminski (2001). Alternatively, schools can create their own reports and summaries.

DIBELS Initial Sound Fluency (ISF) is a standardized, individually administered measure of phonological awareness that assesses a child's ability to recognize and produce the initial sound in an orally presented word (Good, Laimon, Kaminski, & Smith, 2002; Kaminski & Good, 1996). An earlier version of this measure was titled Onset Recognition Fluency (OnRF; Laimon, 1994). The examiner presents four pictures to the child, names each picture, and then asks the child to identify (i.e. point to or say) the picture that begins with the sound produced orally by the examiner. For example, the examiner says, "This is sink, cat, gloves, and hat. Which picture begins with /s/?" and the student points to the correct picture. The child is also asked to produce orally the beginning sound for an orally presented word that matches one of the given pictures. The examiner calculates the amount of time taken to identify/produce the correct sound and converts the score into the number of initial sounds correct in one minute. The ISF measure takes about two minutes to administer and has over 20 alternate forms to monitor progress.

Alternate forms were constructed by first compiling a large pool of appropriate items. Items were generated by selecting words from the PSF word pool that could be represented with a picture. Once an appropriate pool of items was created, items were assigned to benchmark and progress monitoring probes at random and arranged on the probe in a random order. In this manner, all probes were constructed to be a random

sample of items from a common item pool, and thus all probes were equivalent in difficulty through randomization.

DIBELS Phoneme Segmentation Fluency (PSF) is a standardized, individually administered test of phonological awareness (Good, Kaminski, & Smith, 2002; Kaminski & Good, 1996). The PSF measure assesses a student's ability to segment three- and four-phoneme words into their individual phonemes fluently. The PSF measure has been found to be a good predictor of later reading achievement and is intended for use with students from the winter of kindergarten through first grade (Good, Kaminski, & Smith, 2002; Kaminski & Good, 1996). The PSF task is administered by the examiner orally presenting words of three to four phonemes. It requires the student to produce verbally 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. The PSF measure takes about two minutes to administer and has over 20 alternate forms for monitoring progress. In prior research, the two-week, alternate-form reliability for the PSF measure was found to be .88 (Kaminski & Good, 1996).

Equivalent alternate forms of the PSF measure were constructed by first establishing a large pool of appropriate items. The first step in selecting words was to select words that were used in first and second grade reading material from *The educator's word frequency guide* (Zeno, 1995). *The educator's word frequency guide* provides a corpus of words used in written English materials arranged by grade level with an estimate of their relative frequency of occurrence corrected for entropy (symbolized as *U*). Initially, 2687 words were selected where the Grade 1 *U* was greater than or equal to

20 or where the Grade 2 *U* was greater than or equal to 20. Next, words were matched to pronunciations, syllabications, and parts of speech obtained from the *Oxford advanced learner's dictionary*, computer usable version (Mitton, 1986).A more recent version of the computer usable dictionary is available from Hornby, Cowie, and Lewis (1974). Words were excluded from the initial pool if they: (a) were not found in the dictionary, (b) had more than 1 syllable, (c) were identified as proper nouns, (d) included apostrophes, (e) were single phoneme words, (f) were single letter words, (g) had more than 6 phonemes, or (h) were judged to be inappropriate (e.g., die and kill were excluded).

The resulting final PSF item pool consisted of 1346 words and their pronunciations with (a) 46.3% judged to be the easiest words – no r-controlled vowels, no consonant blends, 2 or 3 phonemes; (b) 41.4% judged to be less easy – incorporating one difficulty feature consisting of an r-controlled vowel or a single 2-consonant blend, but not both, no 3-consonant blends, and 2 to 4 phonemes; (c) 2.5% were judged more difficult – incorporating 2 difficulty features, no 3-consonant blends, 2 to 4 phonemes; and (d) 9.8% judged to be most difficult – incorporating 3-consonant blends or 5 phonemes. Alternate form probes were constructed by using a stratified random sampling strategy from the final pool such that probes were comprised of words judged to be: (a) 65% easiest, (b) 30% less easy, (c) 3% more difficult, and (d) 2% most difficult. In this way, each probe contained a range of easier and more difficult items arranged in a random order, and each probe was equivalent in difficulty to all other probes through randomization.

DIBELS Nonsense Word Fluency (NWF) is a standardized, individually administered test of the alphabetic principle – including letter-sound correspondence and

the ability to blend letters into words in which letters represent their most common sounds (Good & Kaminski, 2002; Kaminski & Good, 1996). The student is presented an 8.5″x 11″ sheet of paper with randomly ordered VC and CVC nonsense words (e.g. *sig*, *rav*, *ov*) and asked to produce verbally the individual letter sound of each letter or read the whole nonsense word. For example, if the stimulus word is "vaj" the student could say /v/ /a/ /j/ or say the word /"vaj"/ to obtain a total of three letter-sounds correct. The student is allowed one minute to produce as many letter-sounds as he/she can, and the final score is the number of letter-sounds produced correctly in one minute. Because the measure is fluency based, students receive a higher score if they are phonologically recoding the word and receive a lower score if they are providing letter sounds in isolation. The NWF measure also takes about two minutes to administer and has over 20 alternate forms for monitoring progress.

Equivalent alternate forms were created by first establishing an item pool of all possible eligible nonsense words. In order to be an eligible nonsense word, words had to have each letter associated only with its most frequently occurring sound. Words that were real words or that sounded like inappropriate words were excluded, but words that sounded like real words were not excluded. The final item pool consisted of 1065 words in which: (a) 4% were 2 letter words judged easy – the final consonant was a member of the string of consonants "bcdfghklmnprst" judged to be easier; (b) 1% were 2 letter words judged harder where the final consonant was not an easier consonant; (c) 16% were 3 letter words where only the initial consonant was judged easier; (d) 24% were 3 letter words where only the final consonant was judged easier; (e) 48% were 3 letter words where both consonants were judged easier; and (f) 7% were 3 letter words where both consonants were judged harder. Equivalent alternate form probes were obtained using a

stratified random sampling procedure where: (a) 10% were 2 letter words with easy final consonant, (b) 5% were 2 letter words with harder final consonant, (c) 20% were 3 letter words with easier initial consonant only, (d) 20% were 3 letter words with easier final consonant only; (e) 40% were 3 letter words with easier initial and final consonants; and (f) 5% were 3 letter words with harder initial and final consonants. Harder and easier words were arranged in random order on the probes.

DIBELS Letter Naming Fluency (LNF) is a standardized, individually administered test that provides a measure of risk (Kaminski & Good, 2002). Students are presented with an 8.5″x 11″ sheet of paper with upper-and lower-case letters arranged in a random order and are asked to name as many letters as they can. Students are told if they do not know a letter they will be told the letter. The student is allowed one minute to produce as many letter names as he/she can, and the score is the number of letters named correctly in one minute. Students are considered at risk for difficulty achieving early literacy benchmark goals if they perform below the 20th percentile using local district norms or system-wide norms. Students are considered at some risk if they perform between the 20th and 40th percentile, and are considered at low risk if they perform above the 40th percentile.

Equivalent alternate forms were creating a different random sort of 2 complete upper case alphabets, and 2 complete lower case alphabets for each probe.

Curriculum-Based Measurement of Oral Reading Fluency (CBM ORF) is a standardized procedure to assess a child's accuracy and fluency reading connected text. The measures used in this study are a version of CBM ORF has been published as The Test of Reading Fluency (TORF) (Children's Educational Services, 1987). Subsequent to this study, a set of Oral Reading Fluency passages were created for the DIBELS

assessment system referred to as the DIBELS Oral Reading Fluency (DORF) passages. The TORF is a standardized set of passages and administration procedures designed to (a) identify children who may need further intensive assessment and (b) measure growth in reading skills (Children's Educational Services, 1987, p. #1). Passages were calibrated for each grade level, and student performance is measured by having students read each of three passages aloud for one minute. Words omitted, substituted, and hesitations of more than three seconds are scored as errors. Words self-corrected within three seconds are scored as accurate. The median correct words per minute from the three passages is selected as the oral reading fluency rate.

A series of studies has confirmed the technical adequacy of the TORF. Test-retest reliabilities of elementary students range from .92 to .97; alternate-form reliability of different reading passages drawn from the same level ranged from .89 to .94 (Tindal, Martson, & Deno, 1983). Criterion-related validity studied in eight separate studies in the 1980's reported coefficients ranging from .52 to .91 (Good & Jefferson, 1998).

Woodcock-Johnson Psycho-Educational Battery Total Reading Cluster is a comprehensive measure of reading achievement (Mather & Woodcock, 2001). It includes the subtests, Letter-Word Identification, Reading Fluency, and Passage Comprehension which assess decoding, speed and the ability to comprehend connected text. The median reliability is .93 in the 5 to 19 age range and .94 for adults.

Stanford-Binet Intelligence Scale is an individually administered, norm-referenced measure of general intelligence (Salvia & Ysseldyke, 2001). Two subtests – Verbal Reasoning and Abstract/Visual Reasoning were administered to all participants. Verbal Reasoning involves students answering vocabulary questions and comprehension questions such as, "Why do buildings have fire escapes?" Students are also asked to

explain why pictures presented are absurd and in what way three presented items are alike and the fourth is different. Abstract/Visual Reasoning requires a student to form geometric patterns with cubes, copy designs with blocks or with a pencil, solve matrix-completion problems, and identify what a paper would look like if it was folded and cut a certain way. Test-retest reliability is .91 and .90 for 5-year olds and 8-year olds, respectively. Internal consistency ranged from the high .80s to .90. Concurrent validity with the Wechsler Intelligence Scale for Children-Revised, Full Scale IQ was .83 and .89 with the Mental Processing Composite of the Kaufman Assessment Battery for Children.

Results

The longitudinal results will be reported by measure, with a particular focus on the target time for each measure corresponding to the primary benchmark goal for each measure. For ISF, the target time is the middle of kindergarten when the benchmark goal is 25. For PSF, the target time is the end of kindergarten with a benchmark goal of 35. For NWF, the target time is the middle of first grade with a benchmark goal of 50. Results for the DORF measure are not reported here because the DORF passages were developed subsequent to this study. Information about the development of the DORF passages is available in Good, Kaminski, Smith, and Bratten (2001), Good, Simmons, Kame'enui, Kaminski, and Wallin (2002), and Good and Kaminski (2002).

Initial Sound Fluency

Although data are available for multiple months, December, January and February are particular times of interest. In order for students to be on track for becoming successful readers they need to have established the skill of initial sounds by the winter of

kindergarten (Good, Simmons, & Kame'enui, 2001). Children meeting this goal are on track for meeting the next early literacy goal, which is phonemic awareness.

Table 3 contains the descriptive statistics, one-month alternate-form reliability and concurrent, criterion-related validity for kindergarten ISF. In January of kindergarten the one-month, alternate-form reliability of ISF is .72. In a study by Ditkowsky (2003), the one-week, alternate-form reliability was .72 which may be a more accurate estimate of the reliability of ISF due to the shorter test-retest interval. While these levels of reliability are low with respect to standards for educational decision-making (e.g., Salvia & Ysseldyke, 2001), it is remarkable in a one-minute, early kindergarten, measure – especially one that can be repeated. By repeating the assessment 4 (r = .72) to 6 (r = .62) times, the resulting aggregate is predicted to have a reliability of .91 (Nunnally, 1978).

The median concurrent, criterion-related validity of a single ISF probe with the DIBELS PSF is .48 in winter of kindergarten and .36 with the Woodcock-Johnson Psycho-Educational Battery readiness cluster standard score. ISF was also modestly related to measures of intellectual functioning. In general, ISF displayed lower correlations with intellectual functioning, and the correlations were sometimes non significant.

As presented in Table 4, the median predictive validity of kindergarten ISF with respect to spring of first grade reading on CBM ORF is .38 and is .36 with the Woodcock-Johnson Psycho-Educational Battery total reading cluster standard score. These results are even more remarkable considering the one-year length of the predictive validity interval. Scores on the ISF in the middle of kindergarten significantly predicted student outcomes over a year later - to the end of first grade. Predicting student performance at the end of first grade, based on their performance in the middle of

Table 3

Descriptive Statistics, 1-Month, Alternate-Form Reliability, and Concurrent, Criterion-Related Validity for Kindergarten Initial Sound Fluency

					Correlation with Selected Concurrent Criterion-Related Measures			
Month of Kindergarten	*n*	*M*	*SD*	1-Month, Alternate- Form Reliability	PSF from Concurrent Month	Woodcock-Johnson Readiness Cluster Standard Score	Stanford-Binet Verbal Reasoning Standard Score	Stanford-Binet Abstract/Visual Reasoning Standard Score
December	142	14.33	10.06		.61* (73)	.34* (57)	.41* (123)	.22* (123)
January	142	19.05	12.24	.72* (131)	.48* (142)	.36* (54)	.38* (125)	.23* (125)
February	263	15.78	11.26	.51* (135)	.48* (243)	.45* (61)	.12 (130)	.16 (130)
March	76	20.00	12.98	.63* (000)	.45* (76)	.36* (59)	.12 (59)	.15 (59)
April	78	21.42	12.75	.55* (71)	.45* (78)	.44* (61)	.29* (61)	.30* (61)
May	82	22.54	14.00	.61* (74)	.46* (82)		.26* (68)	.31* (68)

Note. Correlations are based on subjects with pair-wise complete data. The number with pair-wise complete data is reported in parentheses.
*$p < .05$.

Table 4

Predictive, Criterion-Related Validity for Kindergarten Initial Sound Fluency

			May-of-First-Grade Criterion Measure	
Month of Kindergarten	May-of-Kindergarten PSF	December-of-First-Grade NWF	WJ Total Reading Cluster Standard Score	CBMR
December	.35* (66)	.33* (51)	.36* (41)	.45* (51)
January	.45* (62)	.29* (50)	.28 (37)	.30* (50)
February	.34* (75)	.29* (56)	.51* (41)	.38* (56)
March	.34* (71)	.22 (54)	.36* (38)	.34* (53)
April	.46* (74)	.23 (55)	.46* (41)	.26 (55)
May		.32* (60)	.37*(44)	.39* (59)

Note. Correlations are based on subjects with pair-wise complete data. The number with pair-wise complete data is reported in parentheses.
*$p < .05$.

kindergarten, allows educators to identify at-risk students and intervene early to help them get on track for becoming successful readers.

Phoneme Segmentation Fluency

Descriptive statistics, one-month, alternate-form reliability and concurrent, criterion-related validity are presented in Table 5 for PSF. The end of first grade (March, April and May) are of particular interest as the goal assessment period for PSF. Students need to have established phonological awareness skills, by the spring of kindergarten, in order to be on track for later literacy goals (Good, Simmons, & Kame'enui, 2001). The median one-month, alternate-form reliability is .79 in spring of kindergarten. By repeating the assessment three times, the resulting aggregate would be expected to have a

reliability of .92 (Nunnally, 1978). A prior study by Kaminski and Good (1996) found the two-week, alternate-form reliability to be .88 which may be a better estimate of the reliability of PSF due to the shorter test-retest interval.

The median concurrent, criterion-related validity of PSF is .56 with the Woodcock-Johnson Psycho-Educational Battery readiness cluster standard score in spring of kindergarten. The median concurrent, criterion-validity of kindergarten PSF is .38 and .23 with the Stanford Binet Verbal Reasoning and Abstract/Visual Reasoning, respectfully. Low concurrent validity with the Stanford Binet indicates that the measures are assessing two different constructs. This was expected since PSF was developed to measure students' phonological awareness skills and not their reasoning skills.

The median predictive validity of PSF with respect to first grade reading outcomes is reported in Table 6. The predictive validity of PSF in spring of kindergarten with (a) winter of first grade DIBELS NWF is .62, (b) spring of first grade Woodcock-Johnson Psycho-Educational Battery total reading cluster standard score is .63, and (c) spring of first grade CBM ORF is .62. The predictive validity typically increases for PSF from mid kindergarten to the end of kindergarten when it is most valid. Again, these results are remarkable considering the length of the predictive validity interval. We are able to predict how a student at the end of kindergarten will perform on oral reading fluency one year later. This information is critical to for making accurate educational decisions about risk and need for support. Predicting how a student will perform a year later allows educators to intervene early and provide the necessary support for children to become successful readers.

Table 5

Descriptive Statistics, 1-Month, Alternate-Form Reliability, and Concurrent, Criterion-Related Validity for Kindergarten Phoneme Segmentation Fluency

					Concurrent, Criterion-Related Validity		
Month of Kindergarten	*n*	*M*	*SD*	1-Month, Alternate-Form Reliability	Woodcock-Johnson Readiness Cluster Standard Score	Stanford Binet Verbal Reasoning Standard Score	Stanford Binet Abstract/Visual Reasoning Standard Score
December	73	8.59	10.71		.42* (57)	.28* (57)	.23 (57)
January	142	9.67	12.86	.69* (63)	.35* (54)	.26* (125)	.24* (125)
February	246	12.89	12.96	.66* (124)	.54* (61)	.37* (119)	.29* (119)
March	219	17.00	15.25	.74* (196)	.56* (59)	.38* (129)	.23* (129)
April	267	19.27	15.52	.79* (207)	.56* (61)	.34* (128)	.23* (128)
May	232	19.93	15.63	.79* (215)	.54* (66)	.38* (131)	.35* (131)

Note. Correlations are based on subjects with pair-wise complete data. The number with pair-wise complete data is reported in parentheses.
*$p < .05$.

Table 6

Predictive, Criterion-Related Validity for Kindergarten Phoneme Segmentation Fluency

Month of Kindergarten	May-of-Kindergarten NWF	December-of-First-Grade NWF	May-of-First-Grade WJ Total Reading Cluster Standard Score	May-of-First-Grade CBMR
December		.33* (51)	.38* (41)	.35* (51)
January	.49* (63)	.58* (50)	.48* (37)	.50* (50)
February	.39* (125)	.57* (56)	.58* (41)	.53* (56)
March	.38* (134)	.54* (54)	.61* (38)	.50* (53)
April	.37* (141)	.68* (55)	.63* (41)	.63* (55)
May		.62* (60)	.68* (44)	.62* (59)

Note. Correlations are based on subjects with pair-wise complete data. The number with pair-wise complete data is reported in parentheses.
*$p < .05$.

Table 7 contains the descriptive statistics, one-month, alternate-form reliability and concurrent, criterion-related validity for first grade PSF. The median one-month, alternate-form reliability is .67 in first grade. By repeating the assessment three times, the resulting aggregate would be expected to have a reliability of .94 (Nunnally, 1978). In a previous study, the two-week, alternate-form reliability was found to be .60 in first grade (Kaminski & Good, 1996).

The concurrent, criterion-related validity of PSF with the Woodcock-Johnson Psycho-Educational Battery readiness cluster standard score is the highest in October, the beginning of first grade, correlating at .51. In general, the validity of PSF decreases over the course of first grade. The median concurrent validity with the Stanford Binet Verbal Reasoning and Abstract/Visual Reasoning Score is .26 and .20, respectively. Similar to

Table 7

Descriptive Statistics, 1-Month, Alternate-Form Reliability, and Concurrent, Criterion-Related Validity for First-Grade Phoneme Segmentation Fluency

PSF					Concurrent, Criterion-Related Validity		
Month of First Grade	*n*	*M*	*SD*	1-Month, Alternate-Form Reliability	Woodcock-Johnson Readiness Cluster Standard Score	Stanford Binet Verbal Reasoning Standard Score	Stanford Binet Abstract/Visual Reasoning Standard Score
October	82	20.51	16.03		.51* (65)	.26* (82)	.25* (82)
November	90	24.43	16.86	.63* (80)	.42* (64)	.29* (85)	.18 (85)
December	214	23.77	14.74	.70* (87)	.23* (121)	.25* (143)	.15 (143)
January	154	32.73	17.79	.61* (148)	.29* (114)	.33* (136)	.17* (136)
February	297	33.05	15.61	.67* (146)	.31* (121)	.29* (142)	.23* (142)
March	231	36.80	15.80	.65* (219)	.21* (121)	.23* (142)	.22* (142)
April	308	36.90	15.39	.68* (220)	.26* (126)	.28* (147)	.23* (147)
May	242	39.01	14.86	.70* (231)	.19* (125)	.20* (146)	.20* (146)

Note. Correlations are based on subjects with pair-wise complete data. The number with pair-wise complete data is reported in parentheses
*$p < .05$.

kindergarten, low concurrent validity with the Stanford Binet indicates that the measures are assessing two different constructs. This was expected since PSF was developed to measure students' phonological awareness skills and not their reasoning skills.

Table 8 illustrates the predictive validity for first grade PSF with first and second grade reading outcomes. The median predictive validity for PSF is highest in the fall (October and November) of first grade, correlating at (a) .54 with winter of first grade NWF, (b) .54 with winter of first grade CBM, and (c) .58 with spring of second grade Woodcock-Johnson Psycho-Educational Battery total reading cluster standard score. From its time of greatest validity at the end of kindergarten, the validity of PSF appears to decrease during the course of first grade.

When examining these results it is important to consider the length of predictive validity. In this study, we were able to predict how well a first grade student in the fall will perform on oral reading fluency at the end of first grade as well as two academic years later, when the student is in the spring of second grade. This is valuable information to have in order for educators to intervene early and get students on track to meeting later literacy goals.

Nonsense Word Fluency

Table 9 contains the descriptive statistics, one-month, alternate-form reliability and concurrent, criterion-related validity for first grade NWF. The median one-month, alternate-form reliability is .83 in first grade. By repeating the assessment three times, the resulting aggregate would be expected to have a reliability of .94 (Nunnally, 1978).

The median concurrent, criterion-related validity of first grade NWF with the Woodcock-Johnson Psycho-Educational Battery-Revised readiness cluster standard score

Table 8

Predictive, Criterion-Related Validity for First-Grade Phoneme Segmentation Fluency

Month of First Grade	February-of-First-Grade NWF	May-of-First-Grade CBMR	Spring-of-Second-Grade WJ Total Reading Cluster Standard Score	Spring-of-Second-Grade CBMR
October	.53* (74)	.52* (74)	.59* (58)	
November	.55* (82)	.56* (82)	.57* (59)	
December	.44* (197)	.51* (00)	.46* (111)	.34* (53)
January	.43* (147)	.46* (139)	.49* (107)	.32* (51)
February		.22* (214)	.38* (110)	.19 (54)
March		.26* (216)	.31* (114)	.15 (54)
April		.17* (231)	.24* (116)	.08 (57)
May			.20* (114)	.04 (54)

Note. Correlations are based on subjects with pair-wise complete data. The number with pair-wise complete data is reported in parentheses.

is .51 in first grade. Although data are available for multiple months, December, January and February are particular times of interest. Students need to have established alphabetic principle skills by the winter of first grade in order to be on track for meeting later literacy goals (Good, Simmons, & Kame'enui, 2001).

The median concurrent, criterion-related validity of first grade NWF with the Stanford Binet Verbal Reasoning and Abstract/Visual Reasoning is .30 and .32, respectively. Similar to PSF, low concurrent validity of first grade NWF with the Stanford Binet indicates that the measures are assessing two different constructs. This is

Table 9

Descriptive Statistics, 1-Month, Alternate-Form Reliability, and Concurrent, Criterion-Related Validity for First-Grade Nonsense Word Fluency

					Concurrent, Criterion-Related Validity		
Month of First Grade	*n*	*M*	*SD*	1-Month, Alternate-Form Reliability	WJ Readiness Cluster Standard Score	Stanford-Binet Verbal Reasoning Standard Score	Stanford-Binet Abstract/Visual Reasoning Standard Score
October	79	14.29	15.61		.55* (62)	.17 (79)	.21 (79)
November	90	19.24	19.45	.67* (77)	.52* (64)	.28* (85)	.32* (85)
December	214	29.82	23.47	.80* (87)	.37* (121)	.33* (00)	.31* (00)
January	154	35.51	26.03	.83* (148)	.36* (114)	.28* (00)	.31* (00)
February	298	41.22	28.39	.78* (147)	.59* (122)	.36* (143)	.33* (143)
March	233	44.94	30.41	.88* (222)	.51* (123)	.40* (144)	.36* (144)
April	308	48.50	32.64	.87* (222)	.51* (126)	.33* (147)	.37* (147)
May	242	55.41	35.61	.87* (231)	.35* (125)	.27* (146)	.28* (146)

Note. Correlations are based on subjects with pair-wise complete data. The number with pair-wise complete data is reported in parentheses.
*$p < .05$.

expected since NWF was developed to measure the alphabetic principle skill and not a student's reasoning skills.

Table 10 presents the predictive validity of first grade NWF. The median predictive validity of NWF in the middle (December, January, February) of first grade with (a) CBM ORF in May of first grade is .81, (b) CBM ORF in May of second grade is .68, and (c) Woodcock-Johnson Psycho-Educational Battery total reading cluster standard score in May of second grade is .66. With the exception of May of first grade CBM ORF, predictive validity generally increases over the course of first grade.

These results are remarkable considering the length of time of the predictive validity. In this study, we are able to predict how well a first grade student in the winter could perform more than one year later, when the student is in the spring of second grade. Predicting performance at the end of second grade based on their performance in the middle of first grade allows educators to identify at-risk students and intervene early to make sure they are successful readers.

Letter Naming Fluency

Table 11 contains the descriptive statistics, one-month, alternate-form reliability and concurrent, criterion-related validity for kindergarten LNF. The median one-month, alternate-form reliability is .89 in kindergarten. The aggregate of two probes has a reliability of .94 (Nunnally, 1978).

The concurrent, criterion-related validity of kindergarten LNF with the Woodcock-Johnson Psycho-Educational Battery-Revised readiness cluster standard score increases over the year with a median validity of .75 at the end of kindergarten. The median concurrent, criterion-related validity of kindergarten LNF with the Stanford Binet Verbal Reasoning and Abstract/Visual Reasoning is.29 and .24, respectively. Similar to

Table 10

Predictive, Criterion-Related Validity for First-Grade Nonsense Word Fluency

Month of First Grade	May-of-First-Grade CBMR	February-of-Second-Grade CBMR	May-of-Second-Grade WJ Total Reading Cluster Standard Score	May-of-Second-Grade CBMR
October	.71* (70)		.52* (56)	
November	.68* (82)		.54* (59)	
December	.81* (146)	.69* (54)	.64* (111)	.68* (53)
January	.82* (139)	.63* (52)	.66* (107)	.60* (51)
February	.69* (215)	.77* (55)	.72* (111)	.80* (54)
March	.71* (218)	.80* (57)	.67* (116)	.80* (56)
April	.76* (231)	.85* (58)	.77* (116)	.85* (57)
May	.	.72* (55)	.67* (114)	.74* (54)

Note. Correlations are based on subjects with pair-wise complete data. The number with pair-wise complete data is reported in parentheses.
*$p < .05$.

ISF, PSF and NWF, low concurrent validity of kindergarten LNF with the Stanford Binet indicates that the measures are assessing two different constructs. This is expected since LNF is an indicator of risk for meeting later literacy goals and was not developed to measure students' intellectual functioning.

As presented in Table 12, the median predictive validity for kindergarten LNF with: (a) winter of first grade NWF is .71, (b) Woodcock-Johnson Psycho-Educational Battery-Revised readiness cluster standard score is .65, and (c) May of first grade CBM ORF is .71. Again, these results are remarkable when considering the length of the predictive validity interval. From this data we are able to predict how a student in the

Table 11

Descriptive Statistics, 1-Month, Alternate-Form Reliability, and Concurrent, Criterion-Related Validity for Kindergarten Letter Naming Fluency

LNF					Concurrent, Criterion-Related Validity		
Month of Kindergarten	*n*	*M*	*SD*	1-Month, Alternate-Form Reliability	Woodcock-Johnson Readiness Cluster Standard Score	Stanford Binet Verbal Reasoning Standard Score	Stanford Binet Abstract/Visual Reasoning Standard Score
October	74	14.07	13.61			.30* (64)	.17 (64)
November	81	13.98	13.55	.88* (71)		.27* (73)	.19 (73)
December	144	16.15	15.72	.88* (71)	.64* (57)	.29* (125)	.21* (125)
January	142	18.63	15.16	.86* (133)	.66* (54)	.28* (125)	.31* (125)
February	266	22.80	18.22	.90* (135)	.69* (61)	.32* (130)	.27* (130)
March	219	25.71	19.31	.92* (208)	.71* (59)	.30* (129)	.31* (129)
April	267	26.85	19.12	.89* (207)	.75* (61)	.31* (128)	.26* (128)
May	232	31.38	21.10	.90* (215)	.76* (66)	.26* (131)	.23* (131)

Note. Correlations are based on subjects with pair-wise complete data. The number with pair-wise complete data is reported in parentheses.
*$p < .05$.

Table 12

Predictive, Criterion-Related Validity for Kindergarten Letter Naming Fluency

Month of Kindergarten	December-of-First-Grade NWF	May-of-First-Grade Woodcock Johnson Total Reading Cluster Standard Score	May-of-First-Grade CBMR
December	.61* (51)	.44* (41)	.64* (51)
January	.65* (50)	.57* (00)	.72* (50)
February	.71* (56)	.64* (41)	.71* (56)
March	.77* (54)	.69* (38)	.80* (53)
April	.74* (55)	.67* (41)	.72* (55)
May	.72* (60)	.69* (44)	.69* (59)

Note. Correlations are based on subjects with pair-wise complete data. The number with pair-wise complete data is reported in parentheses.
*$p < .05$.

middle of kindergarten will perform on oral reading fluency two years later, at the end of first grade.

Table 13 contains the descriptive statistics, one-month, alternate-form reliability and concurrent, criterion-related validity for first grade LNF. The median one-month, alternate-form reliability is .86 in first grade. The aggregate of two probes would be estimated to have a reliability of .91 (Nunnally, 1978).

The median concurrent, criterion-related validity of first grade LNF with the Woodcock-Johnson Psycho-Educational Battery-Revised readiness cluster standard score is .52 in first grade. The concurrent, criterion-related validity of first grade LNF with the Stanford Binet Verbal Reasoning and Abstract/Visual Reasoning standard score is .27

Table 13

Descriptive Statistics, 1-Month, Alternate-Form Reliability, and Concurrent, Criterion-Related Validity for First-Grade Letter Naming Fluency

					Concurrent, Criterion-Related Validity		
Month of First Grade	*n*	*M*	*SD*	1-Month, Alternate-Form Reliability	Woodcock-Johnson Readiness Cluster Standard Score	Stanford Binet Verbal Reasoning Standard Score	Stanford Binet Abstract/Visual Reasoning Standard Score
October	82	29.95	19.48		.72* (64)	.20 (82)	.18 (82)
November	90	31.08	19.73	.86* (80)	.64* (64)	.27* (85)	.25* (85)
December	215	38.76	20.23	.87* (87)	.45* (121)	.35* (143)	.26* (143)
January	154	45.71	20.84	.80* (148)	.41* (114)	.25* (136)	.20* (136)
February	298	51.06	19.44	.84* (147)	.54* (122)	.30* (143)	.35* (143)
March	233	54.24	22.43	.82* (222)	.52* (123)	.33* (144)	.32* (144)
April	243	56.62	21.79	.87* (222)	.53* (126)	.27* (147)	.37* (147)
May	243	61.12	21.30	.86* (231)	.47* (126)	.28* (146)	.33* (146)

Note. Correlations are based on subjects with pair-wise complete data. The number with pair-wise complete data is reported in parentheses.
*$p < .05$.

and .29, respectively. Similar to kindergarten LNF, low concurrent validity was expected since the measures are assessing two different constructs.

As presented in Table 14, the predictive validity of LNF in December of first grade with (a) spring of first grade CBM ORF is .74, (b) winter of second grade NWF is .68, (c) spring of second grade CBM ORF is .76, and (d) spring of second grade Woodcock-Johnson Psycho-Educational Battery total reading cluster standard score is .61. Once more, these results are especially important considering the length of time for the predictive validity. From this data we are able to predict how a student in the winter of first grade will perform on oral reading fluency one and a half years later, at the end of second grade. Knowing if a student is at-risk for becoming a successful reader in the middle of first grade allows educators to intervene early and make sure the student is on track to meeting later literacy goals.

Discussion

The purpose of this study was to assess the reliability and validity of the DIBELS ISF, PSF, NWF and LNF by looking at longitudinal data across four years and 10 cohorts of children. One limitation of this study was the restriction in geographic location. Data were collected within a single Oregon county and from two elementary schools. Future research is needed with samples from different regions across the nation. Second, the sample sizes of each cohort for whom complete longitudinal data are available were relatively small due to student mobility across the years. Thus, the results of this study must be used with caution when generalizing to other groups of children.

A third limitation is that we were assessing skills that were not necessarily an explicit target within the curriculum. If school identifies students at-risk and teaches

Table 14

Predictive, Criterion-Related Validity of First-Grade Letter Naming Fluency

Month of First Grade	February-of-First-Grade NWF	May-of-First-Grade CBMR	February-of-Second-Grade NWF	May-of-Second-Grade Woodcock Johnson Total Reading Cluster Standard Score	May-of-Second-Grade CBMR
October	.67* (73)	.69* (73)		.58* (58)	
November	.78* (82)	.75* (82)		.59* (59)	
December	.69* (198)	.77* (146)	.61* (54)	.57* (111)	.72* (53)
January	.63* (147)	.76* (139)	.46* (52)	.59* (107)	.48* (51)
February		.72* (215)	.65* (55)	.64* (111)	.79* (54)
March		.74* (218)	.73* (57)	.71* (116)	.83* (56)
April		.74* (231)	.67* (58)	.71* (116)	.79* (57)
May			.66* (55)	.64* (114)	.74* (54)

Note. Correlations are based on subjects with pair-wise complete data. The number with pair-wise complete data is reported in parentheses. *$p < .05$.

targeted skills, obtained predictive validities may decrease from those reported here.

A strength of this study, as well as an interpretive challenge, is the number of reliability and validity coefficients available in this longitudinal study where children were assessed monthly on many of the measures. The availability of multiple coefficients enables an examination of the variability in coefficients. By focusing on median reliability and validity coefficients we can be more confident in the estimate. For each measure we will place particular emphasis on a target timeframe: middle of kindergarten for ISF, end of kindergarten for PSF, middle of first grade for NWF, and the kindergarten year for LNF.

For kindergartners and first graders, all DIBELS measures were moderately reliable at their target times (r = .72) for ISF in January of kindergarten, median r =.79 for PSF at end of kindergarten, median r = .83 for NWF in middle of first grade, median r = .89 for LNF in kindergarten). By repeating the measures 3 (PSF, NWF, LNF) to 4 (ISF) times, the reliability of the resulting aggregate based on the pattern of performance would be quite reliable for all measures (r = .91 to .96). For screening decisions, a reliability of .80 or better is desired, and for important decisions a reliability of .90 is the standard (Salvia & Ysseldyke, 2001). Thus, a single DIBELS probe can serve as an initial screening, but users should be prepared to retest whenever there is a concern about a single probe score. For the most confidence in an estimate of a child's skills, a pattern of performance on repeated assessments across different forms, on different days, and under different conditions should be examined.

The median concurrent validity of the DIBLES measures with the Woodcock-Johnson Broad Reading Cluster was .36 for ISF, .56 for PSF, .51 for NWF, and .75 for

LNF. All DIBELS measures correlated low with the Stanford Binet Verbal and Abstract/Reasoning subtests. This result was expected since the DIBELS were developed to measure early literacy skills and the Stanford Binet was designed to measure reasoning skills. It also indicates that the development of early literacy skills is not highly dependent on verbal skills and intellectual functioning.

The DIBELS PSF measure had moderate validity coefficients at the end of kindergarten and the validity of PSF appeared to decrease over the course of first grade. One possible explanation for the decrease is that children in first grade are beginning to learn more advanced early literacy skills such as alphabetic principle. As children progress through the year the instructional focus shifts from the ability to hear sounds in words, to the ability to associate sounds with letters and use those letter-sounds to read words fluently and accurately. Even though phonological awareness is a fundamental early literacy skill, it is a more basic skill than alphabetic principle and, therefore, not linked as closely to reading outcomes.

Another possible explanation for the decrease is that PSF may display a threshold effect where differences in performance below 35 correct phonemes per minute, or differences between below 35 and above 35 are meaningful, but differences above 35 may not be meaningful. As children move into and through first grade, more and more are scoring above 35 and relations to outcome variables may decrease correspondingly. Thus, in first grade, the essential issue for PSF may be the need to systematically review skills to maintain phonemic awareness skills above 35. For students above 35 on PSF, scores may not be predictive. For the small number of students with continuing deficits in

phonemic awareness in first grade, PSF may be very predictive of difficulty achieving important reading outcomes.

The DIBELS measures display both short-term and long-term predictive validity. In terms of short-term (3-6 months) predictive validity, earlier or more foundational measures are related to later measures: (a) ISF to PSF, median $r = .35$; (b) PSF to NWF, median $r = .62$; (c) NWF to ORF, median $r = .81$. In terms of long-term (at least 12 months) predictive validity, the DIBELS measures at their target times predict both oral reading fluency (ISF median $r = .38$, PSF median $r = .62$, NWF median $r = .69$) and Woodcock Johnson Total Reading Cluster score (ISF median $r = .33$, PSF median $r = .63$, NWF median $r = .66$, LNF median $r = .65$).

Implications for Research and Practice

In this study, two DIBELS measures of phonological awareness were examined, ISF and PSF. Even though both ISF and PSF have adequate technical adequacy, educators may wonder which is a better measure of phonological awareness at which points in time. Determining if a measure is a good indicator of phonological awareness, one needs to look at the reliability, predictive validity and concurrent validity. Results from this study indicate that PSF has higher alternate-form reliability than ISF as well as higher concurrent and predictive validity with respect to important literacy outcomes. These results indicate that PSF is overall a better measure of phonological awareness than ISF. Thus, when both ISF and PSF can be administered, greater confidence should be placed in the PSF measure. However, ISF remains superior at the beginning of kindergarten where PSF has floor effects that preclude its utility. It should be noted that

this study did not provide data for the concurrent validity of ISF or PSF with another assessment of phonological awareness. Future research is needed in that area.

The results presented in this report provide valuable information for identifying students at-risk for reading difficulties. The DIBELS offers educators brief, valid, reliable and repeated measures to assess student's early literacy skills. Knowing how a child performs on the DIBELS measures in kindergarten and first grade strongly predicts their end of first and second grade reading outcomes. Educators can use the DIBELS to identify children, as early as kindergarten, who are at-risk for reading difficulties. Perhaps even more important, DIBELS can provide educators with information to target interventions to core components or early literacy and provide students with support necessary to put them on track for becoming successful readers.

References

Children's Educational Services. (1987). Test of reading fluency. Minneapolis, MN: Author.

Ditkowsky, B. (2002). *Onset recognition computerized assessment system: A validation of measuring the right skills at the right time in the right way*. Unpublished doctoral dissertation, University of Oregon, Eugene.

Good, R.H., Gruba, J., & Kaminski (2001). Best practices in using Dynamic Indicators of Basic Early Literacy Skills (DIBELS) in an outcomes-driven model. In A. Thomas & J. Grimes (Eds.), *Best practices in school psychology IV* (pp. 679-700). Washington, DC: National Association of School Psychologist.

Good, R.H., & Jefferson, G. (1998). *Contemporary perspectives on Curriculum-Based Measurement validity* (pp.61-88)*. New York: Guilford.*

Good, R. H., & Kaminski, R. A. (2002). *Development and Readability of DIBELS Oral Reading Fluency Passages for First through Third Grades* (Technical Report No. 10 No. 9). Eugene, OR: University of Oregon.

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

Good, R.H., Laimon, D., Kaminski, R.A., & Smith, S. (2002). Initial Sound Fluency. In R.H. Good & R.A. Kaminski (Eds.), *Dynamic Indicators of Basic Early Literacy Skills (6th ed.)*. Eugene, OR: Institute for the Development of Educational Achievement. Available: http://dibels.uoregon.edu/

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

Good, R. H., Kaminski, R. A., Smith, S., & Bratten, J. (2001). *Technical Adequacy of Second Grade DIBELS Oral Reading Fluency Passages* (Technical Report No. 8). Eugene, OR: University of Oregon.

Good, R.H., Simmons, D.C., & Kame'enui, E.J. (2001). The importance and decision-making utility of a continuum of fluency-based indicators of foundational reading skills for third-grade high-stakes outcomes. *Scientific Studies of Reading, 5, 257-288.*

Good, R. H., Simmons, D., Kame'enui, E., Kaminski, R. A., & Wallin, J. (2002). *Summary of decision rules for intensive, strategic, and benchmark instructional recommendations in kindergarten through third grade* (Technical Report No. 11). Eugene, OR: University of Oregon.

Good, R.H., Simmons, D.C., & Smith, S.B. (1998). Effective academic interventions in the United States: Evaluating and enhancing the acquisition of early reading skills. *School Psychology Review, 27*(1), 45-56.

Hornby, A. S., Cowie, A. P., & Lewis, J. W. (1974). *Oxford advanced learner's dictionary of current English (Expanded)* (3rd. ed.) [Computer file]. Retrieved May, 2004, from http://ota.ahds.ac.uk/.

Juel, C. (1988). Learning to read and write: A longitudinal study of 54 children from first through fourth grade. Journal *of Educational Psychology, 80*(4), 437-447.

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

Laimon, D.E. (1994). *The effects of a home-based and center-based intervention on at-risk preschool children's early literacy skills*. Unpublished doctoral dissertation, University of Oregon.

McConnell, S., McEvoy, M., Carta, J.J., Greenwood, C.R., Kaminski, R.A., Good, R.H., & Shinn, M. (1998). *Selection of general growth outcomes for children between birth and age eight* (Technical Rep. No. 2). Early Childhood Research Institute on Measuring Growth and Development (ECRI-MGD).

McGill-Franzen, A. (1987). *Failure to learn to read: Formulating a policy problem. Reading* Research Quarterly, 22(4), 475-490.

Mitton, R. (1986). Oxf*ord advanced learner's dictionary (expanded "Computer Usable" version)* (2nd ed.) [Computer file]. Retrieved April, 2000, from ftp://sable.ox.ac.uk:/pub/ota/public/dicts/710/text710.dat

National Reading Panel. (2000*). Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction: Reports of the subgroups.* Bethesda, MD: National Institute of Child Health and Human Development.

Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). New York: McGraw-Hill.

Priest, J.S., McConnell, S.R., Walker, D., Carta, J.J., Kaminski, R.A., McEvoy, M.A., Good, R.H., Greenwood, C.R., & Shinn, M.R. (2001). General growth outcomes for young children: Developing a foundation for continuous progress measurement. *Journal of Early Intervention, 24*(3), 163-180.

Salvia, J., & Ysseldyke, J.E. (2001). *Assessment* (8th ed.). Boston: Houghton Mifflin.

Tindal, G., Marston, D., & Deno, S.L. (1983). *The reliability of direct and repeated measurement (Research Rep. 109).* Minneapolis, MN: University of Minnesota Institute for Research on Learning Disabilities.

U.S. Department of Education (2004). *Reading First.* Retrieved April 22, 2004. Available at http://www.ed.gov/programs/readingfirst/faq.html

Woodcock, R.W., McGrew, K.S., & Mather, N. (2001). *Woodcock Johnson III Tests of Achievement: Examiner's Manual.* Riverside Publishing.

Zeno, S. M. (1995). *The educator's word frequency guide.* Brewster, NY: Touchstone Applied Science Associates.