

**Evaluating Reading Tests for the State of Florida:
Dynamic Indicators of Basic Early Literacy Skills, 6th Edition (DIBELS)
and
Florida Assessments for Instruction in Reading (FAIR)**

Prepared for the
Florida Department of Education
by:

Janet F. Carlson, Ph.D.
Anja Römhild, M.A. (Doctoral Student)
Carina McCormick, M.A. (Doctoral Student)
Katherine Chin, M.S. (Doctoral Student)
Kurt F. Geisinger, Ph.D.
Leslie Shaw, M.A. (Doctoral Student)
Brett P. Foley, M.S. (Doctoral Student)

April 28, 2010

Questions concerning this report can be addressed to:

Kurt F. Geisinger, Ph.D.
Buros Center for Testing
21 Teachers College Hall
University of Nebraska – Lincoln
Lincoln, NE, 68588-0353
kgeisinger2@unl.edu

Executive Summary

The Florida Department of Education asked the Buros Center for Testing to review two reading tests that have been used and may continue to be used in the future in the Florida schools. The Buros Center for Testing, specifically its Buros Institute of Mental Measurements, is without question the best known and largest test evaluation center in the world, with a history of some 75 years working as the “Consumer Reports” of the testing industry.

The two tests reviewed were **DIBELS: Dynamic Indicators of Basic Early Literacy Skills (6th Edition)** and **FAIR: Florida Assessments for Instruction in Reading**. The former instrument is widely available and used in many places and was previously reviewed by Buros in the **Mental Measurements Yearbook**. The latter is an instrument developed by the Florida Department of Education in conjunction with the Florida Center for Reading Research. Although this page represents an executive summary, it obviously would be impossible to summarize some 45 pages of technical text on a page. Both reviews follow the same outline as has been used in the **Mental Measurements Yearbook** series for many years. They open with a description of the instrument and its use, follow with a portrayal of its development, provide a summarization of the technical information by which all educational measures are evaluated, and conclude with both a commentary and a summary. The technical section includes a discussion of standardization and scoring issues, reliability and validity. Evaluative comments are offered primarily in the commentary section, although some evaluative comments occur in other sections in order to place reported information in context.

In accordance with the proposal upon which this description is based, the reviews are not comparative. Each of these instruments has relative strengths and weaknesses. The State of Florida and its competent staff need to consider these reviews in light of the needs it has and decide which test meets which specific purposes.

Table of Contents

Review of <i>Dynamic Indicators of Basic Early Literacy Skills</i> , 6 th edition (DIBELS)	5
Review of <i>Florida Assessments for Instruction in Reading</i> (FAIR)	30

**Review of the
Dynamic Indicators of Basic Early Literacy Skills, 6th Edition (DIBELS)**

Prepared for the
Florida Department of Education

by:

Janet F. Carlson, Ph.D.

Anja Römhild, M.A. (Doctoral Student)

Carina McCormick, M.A. (Doctoral Student)

Katherine Chin, M.S. (Doctoral Student)

Kurt F. Geisinger, Ph.D.

Leslie Shaw, M.A. (Doctoral Student)

Brett P. Foley, M.S. (Doctoral Student)

April 28, 2010

Questions concerning this report can be addressed to:

Kurt F. Geisinger, Ph.D.
Buros Center for Testing
21 Teachers College Hall
University of Nebraska – Lincoln
Lincoln, NE, 68588-0353
kgeisinger2@unl.edu

Description

The Dynamic Indicators of Basic Early Literacy Skills (DIBELS) are a set of brief, individually administered pre-reading and early reading tests designed to assess phonological awareness, alphabetic awareness, and automaticity/fluency in children in Kindergarten (K) through Grade 3. Since its inception, the grade range for which the DIBELS may be used has expanded to include Grades 4 through 6, although most of the related research has centered on K through Grade 3. Assessment in the upper grade range typically involves only oral reading fluency. As well, some use of DIBELS in the pre-Kindergarten range has occurred, but evidence supporting this use is tenuous and not specifically sanctioned by the test developers. Currently in its sixth edition, the next iteration of DIBELS is expected to be available later in 2010 as “DIBELS Next.”

Students complete benchmark assessments three times each year, the results of which are used to evaluate individual student progress in developing early literacy skills. An overview of DIBELS benchmark assessments is presented in Table 1. It includes the following measures: Initial Sound Fluency (ISF; administered in Kindergarten), Letter Naming Fluency (LNF; administered in K – Grade 1), Phoneme Segmentation Fluency (PSF; administered in K – Grade 1), Nonsense Word Fluency (NSF; administered in K – Grade 2), Word Use Fluency (WUF; administered in K – Grade 3), Oral Reading Fluency (ORF; administered in Grades 1 – 3), and Oral Retelling Fluency (RF; administered in Grades 1 – 3). The TORF (Test of Reading Fluency) represents an earlier version of the ORF that is frequently referenced in the DIBELS materials, especially in the technical reports. Under these circumstances, the DIBELS ORF often is referred to as the DORF in order to distinguish it from its predecessor.

Table 1. Overview of Dynamic Indicators of Basic Early Literacy (DIBELS)

	Grade Level			
	K	1	2	3
Initial Sound Fluency (ISF)	x			
Letter Naming Fluency (LNF)	x	x		
Phoneme Segmentation Fluency (PSF)	x	x		
Nonsense Word Fluency (NSF)	x	x	x	
Word Use Fluency (WUF)	x	x	x	x
Oral Reading Fluency (ORF or DORF) / TORF (Test of Reading Fluency)		x	x	x
Oral Retelling Fluency (RF)		x	x	x

Large print versions of DIBELS K-3 stimulus materials are available for downloading from the user site and Braille versions are available at cost to schools. A Spanish version of the DIBELS, known as the Indicadores Dinámicos del Éxito en la Lectura (IDEL), was developed separately from the DIBELS. Although its developers used the DIBELS structure as a springboard for development, the IDEL is not merely a translation of the DIBELS. Review of the IDEL is beyond the scope of the present review; information is available on the DIBELS website. A brief version of the DIBELS, DIBELS Deep, appears to be in development. Information about it is scarce, and no materials are available currently for viewing, downloading, or purchase at the DIBELS user site.

Consistent with the “dynamic” portion of its name, DIBELS was built upon the principles of curriculum-based measurement (CBM) that use *in vivo* methods to gauge student abilities and/or progress. The methodology involves taking brief samples of the

behavior of interest (e.g., reading fluency) frequently, typically within the regular educational setting. Descriptions of CBM generally refer to the idea of inserting a “probe” into a learning context that is authentic (e.g., within a child’s regular classroom). It is tantamount to taking a snapshot during an ongoing adventure.

Chief among the manifold purposes of DIBELS are the identification and progress monitoring of students with underdeveloped basic literacy skills. These children are unlikely to meet state standards for reading when reading is assessed in third grade. DIBELS assessments are intended to be used as screening measures, in that they comprise a “rapid, rough selection process [that is] not followed by further selection procedures” (Anastasi & Urbina, 1997, p. 160). DIBELS provides formative assessment of phonological awareness in young children that may be used to inform diagnostic assessments, yet results may not be used as the sole basis for educational decision making. DIBELS should be used only in conjunction with other diagnostic procedures to establish the presence of reading disabilities. In addition, test results may be used to assess the effectiveness of reading instruction for individual students, as well as for groups of students (e.g., by class or by school).

DIBELS measures are individually administered by teachers or other educational professionals who have been trained in DIBELS administration and scoring. Assessment materials may be downloaded at no charge. Several fee-based data management services are available to help facilitate progress monitoring and score report generation within schools or school districts. These services are not affiliated with DIBELS. Each service has different features, so the extent to which test users can access data beyond the district level (e.g., regional or statewide data) is not known. The test developers offer a variety

of training options, including workshops and an online option. Although training is offered and available, it is not explicitly required, as downloading materials does not require that one submit evidence that one has completed training. Two days of in-person training are regarded as sufficient to prepare test givers to administer and score the test. It seems that two additional days of training prepare test givers to provide training to others seeking to become qualified to administer and score the DIBELS assessments. All DIBELS measures are strictly timed; verbal instructions and directions for administering individual items are precisely scripted. During online training, trainers emphasize standardized administration and caution trainees to “give directions verbatim and time carefully” (Module 1 of online training). They suggest using a stopwatch that keeps time in hundredths of a second, which strikes us as of dubious value, given the ability of most people to operate a stopwatch at this level of precision.

Number of correct responses within a specified period of time (1 minute) comprise the raw scores. The DIBELS benchmark assessments use a database of longitudinal student data to establish minimum performance levels, termed benchmark goals, as well as risk levels for individual DIBELS measures that are considered indicative of future student success or failure in reading. Benchmark assessments are best conceptualized as criterion measures, as the goal is for all students to meet benchmark goals. Locally derived percentile ranks may be developed by individual users to place scores in context, either using 20th and 40th percentiles or other values to index performance deemed appropriate for local needs. Students identified as “at risk” complete progress monitoring assessment throughout the school year. Students classified as at “some risk” should receive targeted instructional support in specific skill areas

where their performance is weakest (J. Wallin, personal communication, April 23, 2010). Progress monitoring assessments include all of the foregoing assessments, except LNF. Test score results provide formative assessments of instructional interventions.

The test developers assert that the DIBELS assessments may be used with the vast majority of students in regular and special education settings, with only a few exceptions. Students who are specifically excluded include students who are deaf, students with speech dysfluency disorders, students learning to read in a language other than English, and students with severe disabilities whose IEPs do not include reading connected text. The Administration and Scoring Guide presents an extensive list of “approved accommodations” that purportedly preserve the integrity of the meaning and interpretation of scores achieved on the assessments. Data-based support for this assertion is not provided. Nevertheless, users are cautioned that results obtained from assessments that are not administered in the standard manner or are not administered using approved accommodations cannot be interpreted and resultant data should not be entered into the web-based data system. Modifying or eliminating time constraints is expressly forbidden as an accommodation. According to the test developers, scores obtained under altered timing or without time limits would not carry similar meaning as scores obtained under standard conditions and would not be interpretable.

Some of the approved accommodations (e.g., providing a quiet setting) comprise commonsense solutions that reduce or eliminate conditions that are unrelated to literacy skills but that may interfere with the assessment process for particular students. Other accommodations incur more risk as far as impacting the meaning and interpretability of test scores. For example, using alternative stimulus pictures in the ISF assessment

depends on the judgment of the examiner both in terms of whether this accommodation is appropriate for a particular child and in terms of selecting the actual stimulus pictures.

Development

DIBELS was developed in an effort to produce an efficient and economical means to gauge student progress towards achieving reading success. The measures align with the reports of both the National Reading Panel (2000) and the National Research Council (1998) that described essential early literacy domains (Hintze, Ryan, & Stoner, 2003; Lundberg, Frost, & Peterson, 1988). It uses techniques of curriculum-based measurement, developed in the 1970s and 1980s, and has been in development and under revision since the late 1980s. DIBELS currently is in the 6th edition, which was published originally in 2002 and revised in 2007. Information from the DIBELS website suggests that release of the 7th edition is imminent.

Theoretical underpinnings of the DIBELS are not described specifically in test materials themselves, such as the Administration and Scoring Guide or the various Technical Reports produced by the test developers. Subtests comprising the DIBELS assess pre-literacy and literacy skills that are generally and widely recognized as essential precursors to reading success. A substantial body of research over about the past 25 years affirms the critical link between phonological awareness, the alphabetic principle, and reading acquisition (Hintze, Ryan, & Stoner, 2003; Roehrig, Petscher, Nettles, Hudson, & Torgesen, 2008). Research also has stressed the importance of several specific skills that represent phonological awareness, some of which are included in the DIBELS (e.g., phoneme segmentation).

Little information is provided in the test materials or on the website about the development of items and stimulus materials (e.g., item writers or sources used to generate items), size of initial item pool, procedures used to review (and, conceivably, revise, delete, or retain) items (e.g., by bias review panels), item tryouts and related analyses (e.g., differential item functioning, Item Response Theory), or pilot testing.

Technical

Standardization

The DIBELS measures were designed to link to specific benchmark goals and to associate specific performance levels with risk categories. Cut-scores for individual performance goals and for the different performance levels were based on longitudinal data collected during the 2001-2002 academic year through the DIBELS Web data system (Good, Wallin, Simmons, Kame'enui, & Kaminiski, 2002). The Web system included data from many participating schools and school districts from the academic year. No evidence could be located that addressed the extent to which participants (i.e., schools and school districts) reflected U.S. demographic characteristics, including geographical location. Good, Wallin, et al. (2002) offer a caveat that the Web data system may have had greater appeal to schools and districts that are “invested in the important beginning reading core areas of phonemic awareness, phonics, and fluency with connected text . . . [and that] . . . endorse the importance of research-based reading curricula and instruction” (p. 2). The acknowledgment of the possibility of differential participation rates appears warranted, but may be incomplete in that it implies that choices and values held by school personnel and leadership are responsible for determining participation. In 2001, widespread availability of Internet access was less

common in school districts than it is today, especially where economic conditions in the surrounding community were strained. Because Web data entry involves a fee and presumably did so in 2001-2002 as well, lack of funds also seems a viable explanation for non-participation by some schools and/or districts.

Overall the total number of students, schools, and districts that participated in specific test administration windows (fall, winter, or spring) and grade levels (Kindergarten through Grade 3) is quite large. Total numbers of students participating ranged between 36,112 and 39,325 in Kindergarten and first grade across the DIBELS measures (with the exception of the LNF measure, where the sample size was 13,221). In Grade 2 and Grade 3, total participants numbered between 10,941 and 16,841.

Descriptive information about participants is limited to the distribution of student numbers within participating districts and schools, and these numbers vary widely, ranging from as low as 1 student per school or district to as high as 4,589 students within a district and 721 students within a school. Sample characteristics such as proportions of students from various racial and ethnic backgrounds, qualifying for free/reduced lunch programs, attending public or private schools, and so forth would be needed in order to evaluate the adequacy of the norming sample as a whole.

Cut-scores associated with specific risk levels were determined on the basis of multiple criteria, which took into consideration the predictive utility of specific cut-scores. The test developers relied upon receiver-operating characteristic curves to optimize false positive/false negative ratios and also considered the overall percentage of students identified within each performance/risk category (i.e., “low risk,” “some risk,” and “at risk”). For the “low risk” category, the cut-score for a particular DIBELS

measure was set such that “the odds in favor of achieving subsequent goals meant that approximately 80% or more of students with the pattern would achieve the goal” (Good, Simmons, Kame’enui, Kaminski, & Wallin, 2002, p. 3). Similarly, the cut-score for the “at risk” category was set so that approximately 20% of students would reach subsequent goals. Finally, the total percentage of students within “at risk” and “some risk” categories at a given test administration window was held to be approximately 20% each. Findings from research conducted by Hintze et al. (2003) indicate that DIBELS cut-scores – notably those for ISF, PSF, and LNF – over-identify students as at risk for related reading difficulties, compared to another widely used measure (Comprehensive Test of Phonological Processing; Wagner, Torgesen, & Rashotte, 1999). Use of alternative cut-scores was suggested as a way to improve classification accuracy – a proposal that was echoed by Nelson (2008), who observed a high false positive rate (67% or higher) for the at-risk cut-score. Results obtained by Roehrig et al. (2008), were generally similar to those of Hintze et al. (2003) and Nelson (2008), but less dramatic as recalibrated cut-scores were found to identify similar proportions of students in risk categories meeting end-of-year benchmarks as original cut-scores. For a screening measure where scores are used primarily to provide formative assessment and to offer a means to monitor progress, over-identification – when it occurs – may not be especially troubling unless it is severe and, therefore, costly. Most educators would agree that it is better to provide services to a child who may not need them than to fail to provide services to a child who does need them.

Some aspects of the cut-score determination process remain unclear. For example, it is generally not known which DIBELS measures were used as criterion

benchmark goals in determining the cut-scores for each DIBELS measure at each of the relevant test administration windows. Technical Report No. 11 (Good, Simmons, et al., 2002) includes tables meant to assist in the formation of instructional interventions based on various patterns of performance on the DIBELS measures. The information in these tables is not exhaustive and includes only some of the possible DIBELS measures that could serve as criterion measures. No rationale is provided as to why these were included and others not. Summary information is provided concerning both overall student sample sizes per grade level/administration window as well as sample size distributions across school districts and schools. However, because information from these samples was not used to generate actual norms, appropriateness of the sample sizes cannot be evaluated, per se.

Reliability

Reliability is addressed in the form of alternate-form reliability coefficients and, to a lesser extent, test-retest reliability coefficients for individual DIBELS measures. These values are reported in the Administration and Scoring Guide (Good & Kaminski, 2002), Technical Report No. 9 (Good, Wallin, et al., 2002), as well as a document currently on the DIBELS website entitled “Assessment Committee Analysis of Reading Assessment Measures” (https://dibels.uoregon.edu/techreports/dibels_5th_ed.pdf).

Evidence supporting reliability is incomplete, as information is not provided for all relevant grades (e.g., no reliability information for Grade 3 is presented) and little or no information is provided concerning sample characteristics or specific calculations that were implemented. Coefficients and sample sizes are presented as ranges, making evaluation of individual measures imprecise. In general, reliability coefficients trend in

the moderate to moderately high ranges. With the largest sample referenced at 231, sample sizes appear modest, at best.

Curiously, the test developers describe most alternate-forms reliability evidence using data collected from multiple administrations (between 5 and 7) spread throughout an academic year, each of which used a different form of the test. In the absence of further information about the actual calculations, it appears that these coefficients reflect test score stability across time as well as across forms of the test. Unfortunately, test-retest and alternate-forms reliability are confounded under this methodology. Nevertheless, reliability evidence is reported in terms of alternate-forms reliability for all measures and as test-retest reliability for the ORF (TORF and DORF). This information is summarized in the following paragraph, the previously noted dearth of specific information that would help contextualize these values notwithstanding.

The alternate-form reliability coefficient for the ISF was reported as .72, following administration in January of the Kindergarten year. Repeating the administration four times increased the average coefficient to .91. Identical data were reported for Onset Recognition Fluency (OnRF), the precursor of the ISF. Data collected at five times over the course of an academic year using alternate forms yielded coefficients ranging from .51 to .73, with a median value of .61. The largest sample size reported was 135. It appears that some sample sizes were not reported. Reliability coefficients for the PSF were reported as .88 and .79, based on a 2-week and 1-month testing interval, respectively, in May of the Kindergarten year. For Kindergarten, data collected at five times over the course of an academic year using alternate forms yielded coefficients ranging from .66 to .79, with a median value of .74 and a reported sample

size range from 63 to 215. For Grade 1, data collected at five times over the course of an academic year using alternate forms yielded coefficients ranging from .60 to .70, with a median value of .67 and a reported sample size range from 80 to 231. The alternate-form reliability coefficient for the NWF was reported as .83, based on a 1-month testing interval during January of the Grade 1 year. Data collected at seven times over the course of an academic year using alternate forms yielded coefficients ranging from .67 to .88, with a median value of .83 and a reported sample size range from 77 to 231. The coefficient for the LNF was reported as .88, based on a 1-month testing interval of the Kindergarten year. For Kindergarten, data collected at seven times over the course of an academic year using alternate forms yielded coefficients ranging from .86 to .92, with a median value of .89 and a reported sample size range from 71 to 215. For Grade 1, data collected at seven times over the course of an academic year using alternate forms yielded coefficients ranging from .80 to .87, with a median value of .86 and a reported sample size range from 80 to 231.

Test-retest reliability coefficients for the TORF and the DORF are very high, generally at the .90 level or higher, with median values near the mid-.90 range (cf. Good, Wallin, et al., 2002). Test-retest intervals are not specified.

Validity

The primary use of DIBELS results is to indicate which students may be at risk for reading below grade level. Results also are used to formulate interventions for students identified as at risk for reading difficulties. Progress monitoring and assessment of educational intervention comprise other potential uses of test results. Validity

evidence offered explicitly by the distributors of DIBELS centers primarily on its use as a screening measure.

Validation evidence attributable to the developers of DIBELS is quite limited. During online training, coverage of technical properties is given only cursory treatment. In one of the ten modules that comprise online training, a single slide is presented with a column of validity coefficients. The trainer makes a brief comment that affirms the technical robustness of DIBELS. However, in the absence of additional information pertaining to the participants and methodology, these data cannot be meaningfully interpreted.

Good, Wallin, et al. (2002) report validity information for the ISF, PSF, NWF, and LNF measures by providing concurrent and predictive evidence. Concurrent evidence was derived from comparisons of performance by Kindergarten and Grade 1 students on the individual DIBELS measures and the Readiness Cluster of the Woodcock-Johnson Psycho-Educational Battery. Resulting correlation coefficients were .36, .54, .36 (in January of Grade 1; in February of Grade 1 the correlation coefficient was .59), and .70, respectively. Individual DIBELS measures were used to predict later performance on the ORF measure in Grade 1. Corresponding correlation coefficients were .45, .62, .82, and .71, respectively. Performance on individual DIBELS measures also were compared to subsequent performance on the Total Reading Cluster score of the Woodcock-Johnson Psycho-Educational Battery, all presumably in Grade 1. Corresponding correlation coefficients were .36, .68, .66, and .65, respectively. A few additional coefficients of similar magnitude were reported for other pairings (e.g., ISF with PSF during January of the Kindergarten year yielded a coefficient of .48).

Criterion evidence supporting validity of the ORF measure was assessed in a series of eight studies that were conducted in the 1980s and reported by Good and Jefferson (1998). The studies used the TORF and reported correlation coefficients in the moderate to high ranges, from .52 to .91. The DORF was “developed to be consistent in readability, reliability, and validity to the TORF passages” (Good, Wallin, et al., 2002, p. 9). Concurrent validity coefficients between TORF and DORF were very high, ranging from .92 to .96.

The bulk of validity evidence must be sought out in studies undertaken by researchers who apparently are not affiliated with the University of Oregon or the DIBELS test authors or developers. A partial list of empirical studies related to the psychometric properties is available on the DIBELS website, in no particular order and without annotation. The ever-widening use of DIBELS seems to have prompted a great deal of related research, such that there is now an extensive body of research, including scores of doctoral dissertations. Much of the available validity evidence concerns the degree to which DIBELS scores correspond with scores from other similar measures and the extent to which DIBELS scores can predict reading performance on reading or reading-related tests.

Overall, correlation coefficients reported in numerous external studies are similar in magnitude to those reported by Good, Wallin, et al. (2002) – generally moderate levels with a fairly wide range of values. For example, Hintze et al. (2003) compared scores obtained on the DIBELS measures with those achieved on the Comprehensive Test of Phonological Processing (CTOPP; Wagner et al., 1999). Phonological Awareness (PA) and Phonological Memory (PM) composite scores from the CTOPP were moderately

correlated with DIBELS measures (ISF, PSF, and LNF). Coefficients ranged from .53 to .60 and from .39 to .52, respectively. The authors conclude that DIBELS and CTOPP assess similar constructs. Burke, Hagan-Burke, Kwok, and Parker (2009) used a longitudinal framework to predict DIBELS ORF and other reading measures in Grades 1 and 2. Some predictive relationships were noted between DIBELS measures administered in Kindergarten and subsequent performance on ORF and other measures administered in Grades 1 and 2, but coefficients rarely exceeded .50. Pressley, Hilden, and Shankland (2005) reported results of a regression analysis, regressing TerraNova (CTB/McGraw-Hill, 2002) scores on DIBELS ORF scores. The resulting coefficient was characterized as relatively low ($r = .446$; $R^2 = 20\%$). However, considering the brevity of the DIBELS ORF, we would probably characterize the coefficient as moderate. Pressley et al. (2005) also questioned the probity of assessing reading fluency by using a test of decoding speed – to the near exclusion of comprehension (correlation coefficients of .16 and .32 were noted between ORF scores and oral retelling of idea units).

Comparisons of ORF scores to state-mandated reading achievement scores for Grade 3 yielded moderately high coefficients of .67 (Good, Simmons, & Kame'enui, 2001), .70 (Buck & Torgesen, 2003), and ranging from .73 to .80 (Barger, 2003; Shaw & Shaw, 2002; Wilson, 2005, all as cited in Riedel, 2007). Research by Riedel (2007) used forward stepwise logistic regression to demonstrate that the ORF was the best predictor of Grades 1 and 2 reading comprehension evaluated at the end of the academic year. Coefficients were .77 and .80 for ORF assessed at the middle and end of Grade 1, respectively, and .67 and .71 for ORF assessed at the middle and end of Grade 2, respectively. Coefficients for the other DIBELS measures administered at the same times

were appreciably lower, ranging from .53 to .69 for Grade 1 and from .54 to .67 for Grade 2.

Research by Roehrig et al. (2008) examined predictive and concurrent evidence related to the ORF scores in Florida, using the Reading portion of the Florida Comprehensive Assessment – Sunshine State Standards (FCAT) and the Stanford Achievement Test 10th edition (SAT-10) scores in Grade 3. Overall, correlation coefficients ranged from .66 to .71 and from .68 to .71 for the FCAT and SAT-10, respectively. The researchers also investigated whether the ORF appeared to demonstrate predictive bias across demographic subgroups, using large calibration ($n = 16,539$) and cross-validation ($n = 16,908$) samples of Florida third graders, stratified to match several demographic features of the state’s population (i.e., race/ethnicity, English Language Learners, students with disabilities, students in gifted programs; 75% of the sample were eligible for free or reduced-price lunch). Logistic regression analysis indicated no significant interactions between ORF scores and membership in the focal groups.

Several authors (Hintze et al, 2003; Riedel, 2007) have observed that the measures comprising DIBELS were developed to reflect reading skills emphasized by the National Reading Panel report (National Institute of Child Health and Human Development, 2000), including the alphabetic principle, phonemic awareness, and fluency. Empirical support for the theoretical foundation of DIBELS has been provided in various reviews and meta-analyses of the research literature (e.g., Ehri, Nunes, Stahl, & Willows, 2001; Ehri, Nunes, Willows, et al., 2001; Rayner, Foorman, Perfetti, Pesetsky, & Seidenberg, 2001; Snow, Burns, & Griffin, 1998). However, we could not locate specific information about precisely who wrote test items or developed reading

passages. As well, information about item review processes (e.g., how many items were initially developed and what retention guidelines were followed) was not evident.

Whether item review panels were convened or used in any way (e.g., to flag items that are potentially biased) is unknown.

Commentary

By design, the DIBELS measures provide formative information and offer a variety of tools and tasks intended for use between more formal assessments – that is, for instructional purposes. The development of tasks appropriate for such ongoing progress monitoring speaks directly to this issue and is responsive to the needs of individuals in the classroom – students and teachers. That so many of the DIBELS materials may be downloaded for free makes them attractive to a wide audience. Online availability of DIBELS materials probably comprises both a plus and a minus. Ready access to test materials is convenient and appealing to many educational professionals and virtually assures widespread use. The website includes a warning to parents advising them not to access test materials. Although there is no easy way to monitor who downloads materials, we believe that the warning is well stated and that the ultimate goals of the DIBELS measures may be reasonably well preserved by the effort to limit access. At the same time, we wonder about the unintended consequences (e.g., on test score validity) that unfettered access to test materials may produce.

Oral reading fluency is recognized as an important correlate of reading comprehension. Of all the DIBELS measures, the strongest empirical support surrounds the ORF. Other DIBELS measures have fared less well, when it comes to establishing an

empirical base that follows from the theoretical one. Overall, results have been mixed and statistically significant coefficients, when observed, have been modest in size.

All DIBELS measures are “fluency based,” where fluency is operationalized as speed of accurate performance or responding. Some evidence supports this decision for certain subtests (e.g., ORF), but many writers have expressed ongoing concern about whether speed comes at the expense of other important aspects of reading such as comprehension. We share these concerns and also wonder whether the need for speed is communicated adequately to test takers and whether test administrators can reliably time 3 seconds (the typical length of time after which an item is passed over and the next item administered). We echo concerns expressed by Samuels (2007) about whether fluency, as operationalized throughout the DIBELS, is an appropriate descriptor. As stated by Samuels, “it is the simultaneity of decoding and comprehension that is the essential characteristic of reading fluency....Secondary characteristics of fluency such as speed, accuracy, and expression are indicators, but not the essential characteristics” (p. 564).

The absence of a Technical Manual constitutes a grave concern. It is our understanding that the next iteration of DIBELS eventually will consolidate technical information in an Assessment Guide and a Technical Manual. More complete documentation that not only consolidates but provides much needed contextual detail concerning psychometric information and evidence is sorely needed. In addition, DIBELS training requirements are vague, although it is difficult to imagine learning the ins and outs of test administration and scoring without some type of formal training. Administration is replete with instances where test administration or scoring of items depends on a subjective decision by the examiner. Several DIBELS measures require

that scoring be completed “on the fly,” which may be problematic as far as accuracy and also may disrupt rapport, diminish motivation, or raise anxiety. Despite the trainers’ calls for adherence to standardized procedures in administration, there is plenty of room for slippage. In this context, the absence of evidence supporting inter-rater reliability is troubling. As an individually administered test, administration and scoring of DIBELS depends greatly upon the skill of the examiner including the precision with which timing of 3 seconds occurs. We would encourage greater specificity in DIBELS materials regarding the essential nature of training.

We note that the test developers evaluate test-retest reliability using intervals that collectively span an entire academic year. However, DIBELS is meant to assess skill development in reading. Measures taken at different points in time separated by many weeks or months would be expected to reflect growth (i.e., increased performance) rather than stasis (i.e., consistent performance). We would encourage additional investigations of test-retest reliability using shorter time intervals and the same test form, in order to produce clear evidence of the DIBELS short-term temporal stability.

It is not apparent how decisions to accommodate students’ needs are rendered. To us, this decision seems to lie largely or entirely within the administrator’s discretion, as long as the accommodation is on the list of approved accommodations. It also appears that it is up to the test giver whether to regard a child’s articulation difficulties as acceptable for that child based on the administrator’s on-the-spot judgment or, possibly, prior knowledge of that child’s speech patterns. Data to support the use of “approved accommodations” are most certainly needed. It appears that the test developers used a subjective approach to determine which accommodations they thought would not affect

the interpretations or utility of resultant scores. Ultimately, however, the impact of accommodations is an empirical question, the answer to which requires the accumulation of data showing how score interpretations remain valid after given accommodations have been implemented.

The work of Riedel (2007) and Roehrig et al. (2008) notwithstanding, further empirical evidence is needed to establish the utility of the DIBELS measures (especially PSF and NWF) using “students with a broad range of reading abilities in high poverty urban settings” (Riedel, 2007, p. 550). Data reported from 2001-2002 include system wide percentile ranks for DIBELS benchmark assessment (Good, Wallin, et al., 2002), available at https://dibels.uoregon.edu/techreports/DIBELS_Percentiles.pdf, effectively constitute the norms for the DIBELS. These data need to be updated and need to specify demographic characteristics of the sample.

Summary

The DIBELS represent a reading measure currently in wide use throughout the nation for students in Grades K-6, with primary use in Grades K-3. Its components are readily available for free downloading at the DIBELS website and are quick to administer. Scoring is complicated and, in places, subjective. The main purpose of DIBELS is to provide formative assessments that can be used to guide instruction that ultimately improves reading education for all students. Performance on the DIBELS measures is used to identify students who are considered “at risk”, at “some risk”, or at “low risk” for failing to meet grade level reading standards. Although the DIBELS measures have been available for many years, documentation remains weak. Psychometric evidence is scattered throughout the professional literature making it

difficult for prospective users to evaluate the test's suitability for their intended uses. A considerable body of research related to DIBELS offers a moderate amount of evidence that supports its use, more so for the oral reading fluency measure than for the other measures. However, some authors argue that the DIBELS measures emphasize speed (fluency) at the cost of sufficient emphasis or information on reading comprehension. Some of its shortcomings may be resolved soon, as the next iteration of DIBELS (DIBELS Next) is slated to be released within the present calendar year and a "comprehensive Technical Manual" is to follow next year (J. Wallin, personal communication, April 23, 2010).

References

- Anastasi, A., & Urbina, S. (1997). *Psychological testing*. New York: Macmillan.
- Burke, M. D., Hagan-Burke, S., Kwok, O., & Parker, R. (2009). Predictive validity of early literacy indicators from the middle of Kindergarten to second grade. *The Journal of Special Education, 42*, 209-226.
- CTB/McGraw-Hill. (2002). *TerraNova*. Monterey, CA: CTB/McGraw-Hill.
- Ehri, L. C., Nunes, S. R., Stahl, S. A., & Willows, D. M. (2001). Systematic phonics instruction helps students learn to read: Evidence from the National Reading Panel's meta-analysis. *Review of Educational Research, 71*, 393-447.
- Ehri, L. C., Nunes, S. R., Willows, D. M., Schuster, B. V., Yaghoub-Zadeh, Z., & Shanahan, T. (2001). Phonemic awareness instruction helps children to read: Evidence from the National Reading Panel's meta-analysis. *Reading Research Quarterly, 36*, 250-287.
- Good, R. H., & Jefferson, G. (1998). Contemporary perspectives on Curriculum-Based Measurement validity. In M. R. Shinn (Ed.), *Advanced applications of Curriculum-Based Measurement* (pp. 61-88). New York: Guilford.
- Good, R. H., & Kaminski, R. A. (Eds.). (2002). *Dynamic Indicators of Basic Early Literacy Skills* (6th ed.). Eugene, OR: Institute for the Development of Achievement.
- 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. S., Kame'enui, E. J., Kaminski, R. A., & Wallin, J. (2002). *Summary of decisions rules for intensive, strategic, and benchmark instructional recommendations in Kindergarten through third grade* (Tech. Rep. No. 11). Eugene, OR: University of Oregon.
- Good, R. H., Wallin, J., Simmons, D. C., Kame'enui, E. J., & Kaminski, R. A. (2002). *System-wide percentile ranks for DIBELS benchmark assessment* (Tech. Rep. No. 9). Eugene, OR: University of Oregon.
- Hintze, J. M., Ryan, A. L., & Stoner, G. (2003). Concurrent validity and diagnostic accuracy of the Dynamic Indicators of Basic Early Literacy Skills and the Comprehensive Test of Phonological Processing, *School Psychology Review*, 32, 541-546.
- Lundberg, I., Frost, J., & Peterson, O. P. (1988). Effects of an extensive program for stimulating phonological awareness in preschool children. *Reading Research Quarterly*, 23, 263-284.
- 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*. Washington, DC: National Institute of Child Health and Human Development.
- National Research Council (1998). *Preventing reading difficulties in young children*. Washington, DC: National Academy Press.
- Nelson, J. M. (2008). Beyond correlational analysis of the Dynamic Indicators of Basic Early Literacy Skills (DIBELS): A classification validity study. *School Psychology Quarterly*, 23, 542-552.

- Pressley, M., Hilden, K., & Shankland, R. (2005). *An evaluation of end-grade-3 Dynamic Indicators of Basic Early Literacy Skills (DIBELS): Speed reading without comprehension, predicting little* (Tech. Rep.). East Lansing, MI: Michigan State University, Literacy Achievement Research Center.
- Rayner, K., Foorman, B. R., Perfetti, C. A., Pesetsky, D., & Seidenberg, M. S. (2001). How psychological science informs the teaching of reading. *Psychological Science in the Public Interest*, 2, 31-74.
- Riedel, B. W. (2007). The relation between DIBELS, reading comprehension, and vocabulary in urban first-grade students. *Reading Research Quarterly*, 42, 546-567.
- Samuels, S. J. (2007). The DIBELS tests: Is speed of barking at print what we mean by reading fluency? [Commentary]. *Reading Research Quarterly*, 42, 563-566.
- Snow, C. E., Burns, S. M., & Griffin, P. (Eds.). (1998). Preventing reading difficulties in young children. Washington, DC: National Academy Press.
- Wagner, R. K., Torgesen, J. K., & Rashotte, C. A. (1999). *Comprehensive Test of Phonological Processing*. Austin, TX: PRO-ED.

**Review of the
Florida Assessments for Instruction in Reading (FAIR)**

Prepared for the
Florida Department of Education
by:

Janet F. Carlson, Ph.D.
Carina McCormick, M.A. (Doctoral Student)
Anja Römhild, M.A. (Doctoral Student)
Katherine Chin, M.S. (Doctoral Student)
Kurt F. Geisinger, Ph.D.
Leslie Shaw, M.A. (Doctoral Student)
Brett P. Foley, M.S. (Doctoral Student)

April 28, 2010

Questions concerning this report can be addressed to:

Kurt F. Geisinger, Ph.D.
Buros Center for Testing
21 Teachers College Hall
University of Nebraska – Lincoln
Lincoln, NE, 68588-0353
kgeisinger2@unl.edu

Description

The Florida Assessments for Instruction in Reading (FAIR) are a set of brief, individually administered literacy tests designed to assess early reading or pre-reading skills (K-2) and reading skills (Grades 3-12). The assessments are administered three times a year, either by the classroom teacher (K-2) or through a computer adaptive testing protocol (Grades 3-12). Teachers in Grades K-2 may enter data using a computer program that allows student data to be submitted through Florida's Progress Monitoring and Reporting Network (PMRN), a "state-wide web-based system that provides reports of student progress at the individual student, classroom, school, district, and state level" (FAIR K-2 Administration Manual, p. 4).

By design, classroom teachers administer the FAIR in Grades K-2, following training that is provided by the Florida State Department of Education. Teachers complete approximately one and a half days of grade-specific training provided by a master trainer. Teachers receive training in assessment administration, data entry, the electronic scoring tool, and instructional implications. The electronic scoring tool also helps guide administration indicating, for example, which test to administer next. Although assessments administered in K-2 differ from those administered in Grades 3-12 in that they are not "adaptive" technically, certain decisions regarding which passage to administer are based on earlier performance. Likewise, students who commit errors on more than 10% of the words in a given reading comprehension passage are interrupted and the next lower level passage is administered. Training for Grades 3-12 is completed in one day and covers online test administration, instructional implications, and the PMRN system.

The FAIR is intended to be used with the majority of students in general and special education settings in the state of Florida. The FAIR Administration Manuals (K-2 and Grades 3-12) present extensive lists of “approved accommodations” grouped into five categories depending on what aspect of standard testing is altered: Presentation, Responding, Scheduling, Setting or Tester, and Assistive Devices, while allowing for the possibility that the FAIR may not be an appropriate assessment for some students with disabilities or for some English Language Learners. Using any of approved accommodations purportedly preserves the integrity of the meaning and interpretation of scores achieved on the assessments. However, data-based support for this assertion is not provided. Nevertheless, users are cautioned that results obtained from assessments that are not administered in the standard manner or that are not administered using approved accommodations cannot be interpreted. Some of the approved accommodations (e.g., providing a special room with minimal distractions, complete quiet, or specialized lighting) comprise commonsense solutions that reduce or eliminate conditions that are unrelated to literacy skills but that may interfere with the assessment process for particular students.

Test components vary by grade levels in the early grades and also by assessment periods. At all grade levels, all students complete a Broad Screen/Progress Monitoring measure (hereafter, Broad Screen), the results of which are used to calculate success probabilities (Probability of Reading Success, or PRS, in K-2 and FCAT Success Probability, or FSP, in Grades 3-12). FAIR K-2 includes two additional inventories of tests – the Broad Diagnostic Inventory and the Targeted Diagnostic Inventory. FAIR Grades 3-12 includes the Targeted Diagnostic Inventory. The latter inventories are used

to gather more specific information related to student instructional needs. Tasks used for Ongoing Progress Monitoring in K-2 as well as tasks in the Grades 3-12 Informal Diagnostic Tool Kit are highly similar or even identical to components of the FAIR, but are not used as tests, per se. Tasks comprising the Informal Diagnostic/Progress Monitoring Assessment for Grades 3-12 Tool Kit are used primarily by intervention teachers to monitor progress very closely and to inform everyday instructional decision making.

In light of the foregoing, we restrict the remainder of this review to the formal measures comprising the FAIR, which are administered thrice a year during prescribed assessment periods. Each assessment period is 34 instructional days in length. Assessment Periods 1, 2, and 3 occur during the fall, winter, and spring, respectively, and include Instructional Days 6 – 40, 66 – 100, and 136 – 170, respectively. Teachers may administer the assessments on any instructional day that falls within the range associated with the corresponding assessment period. Table 1 summarizes the FAIR components and tests within the components.

Table 1. Overview of the Florida Assessments for Instruction in Reading (FAIR)

	Grade Level			
	K	1	2	3-12
Broad Screen/Progress Monitoring				
Letter Names and Sounds	x			
Phonemic Awareness	x			
Word Reading	x	x	x	
Reading Comprehension				x
Broad Diagnostic Inventory				
Letter Sound Knowledge	x			
Listening Comprehension	x	x ^b	x ^b	
Vocabulary	x	x	x	
Reading Comprehension	x ^a	x	x	
Spelling			x	
Targeted Diagnostic Inventory				
Print Awareness	x			
Letter Name and Sound Knowledge	x	x		
Phoneme Blending	x	x		
Phoneme Deletion	x	x	x	
Letter Sound Connections	x			
Word Building	x	x	x	
Multisyllabic Word Reading			x	
Maze				x
Word Analysis				x

^a Assessed with readers only, at Assessment Period 3.

^b Assessed with non-readers.

Kindergarten-level tests comprising the Broad Screen component consist of Letter Names and Sounds, Phonemic Awareness, and Word Reading. The resulting PRS score is used to distinguish levels of literacy competence – red success zone, yellow success zone, and green success zone, that reflect, respectively, < 15%, 16% - 84%, and > 85% chance of performing successfully on the end-of-year reading assessment. The PRS

score, which is expressed as a percent probability, is derived from a calculation that predicts the probability of the student scoring at or above the 40th percentile on the end-of-year test (i.e., SESAT Word Reading), consistent with standard federal reporting requirements. Tests in the Broad Diagnostic Inventory are used to provide specific information related to individual students' needs in reading instruction. These tests include Letter Sound Knowledge (optional test for students with PRS > 85%), Listening Comprehension, and Vocabulary; Comprehension Placement Word List and Reading Comprehension are administered to students with PRS > 85% during the final assessment period. Tests in the Targeted Diagnostic Inventory are administered to students with PRS < 85% and include tasks linked to Florida's Sunshine State Standards and key instructional reading skills. These tests include Print Awareness (an optional test), Letter Name and Sound Knowledge, Phoneme Blending, Phoneme Deletion (2 variants), Letter Sound Connections (2 variants), and Word Building (3 variants).

In Grade 1, the Broad Screen includes one test – Word Reading. As with Kindergarten, the Broad Screen yields a PRS score that is used to distinguish levels of literacy competence (red, yellow, or green zone) associated with the probability performing successfully on the end-of-year reading assessment – SAT-10 Reading Comprehension. Tests in the Broad Diagnostic Inventory include Comprehension Placement Word List, Reading Comprehension (if performance on placement list so indicates, Listening Comprehension may be used instead), and Vocabulary. Tests in the Targeted Diagnostic Inventory (for students with PRS < 85%) include Letter Sound Knowledge, Phoneme Blending, Phoneme Deletion (2 variants), and Word Building (4 variants).

In Grade 2, the Broad Screen includes also includes the Word Reading test. As with Kindergarten and Grade 1, the Broad Screen yields a PRS score that is used to distinguish levels of literacy competence (red, yellow, or green zone) associated with the probability performing successfully on the end-of-year reading assessment – SAT-10 Reading Comprehension. Tests in the Broad Diagnostic Inventory include Comprehension Placement Calculation Procedure, Reading Comprehension (2 variants), Vocabulary (optional for students with PRS > 85%), and Spelling (which is group administered). Tests in the Targeted Diagnostic Inventory (for students with PRS < 85%) include Phoneme Deletion (2 variants), Word Building (3 variants), and Multisyllabic Word Reading.

In Grades 3-12, the Broad Screen includes the Reading Comprehension test, which “intentionally mirrors” (FAIR Grades 3-12 Administration Manual, p. 85) the Florida Comprehensive Assessment Test (FCAT). Performance on this measure is combined with prior-year FCAT performance to generate an FCAT Success Probability (FSP) score that is used to distinguish levels of literacy competence (red, yellow, or green zone) based on the probability of scoring at an acceptable level (i.e., Level 3) on the current year’s FCAT. Tests in the Targeted Diagnostic Inventory are required for students with an FSP < 85% and optional for students with an FSP > 85%. The tests include Maze and Word Analysis. All measures may be used as tasks for ongoing progress monitoring purposes within classroom settings.

In addition to the PRS scores (K-2) and FSP scores (Grades 3-12), a variety of other types of scores are available. For Grades 3-12, the Broad Screen measure itself can be used to obtain several other types of scores in addition to the FSP, including Percentile

Ranks, Standard Scores, Developmental Ability Scores, Reading Comprehension Ability Ranges, Cluster Area Scores, Lexile Measures, and Lexile Ranges. The Developmental Ability Scores were constructed using complex pretesting and equating procedures to produce a useful scale that can show growth across assessment periods within a school year, as well as growth across school years. Developmental Ability Scores are also available for the Reading Comprehension task in Ongoing Progress Monitoring and for Word Analysis in Grades 3-12. Adjusted Fluency Scores are available for Grades 1-5 and are reported in the PMRN for Ongoing Progress Monitoring – Oral Reading Fluency. Adjusted fluency equates difficulty across reading passages, thereby allowing a student's performance to be compared throughout the school year. The standard error of measurement around Developmental Ability Scores represents the Ability Range Scores. Cluster Area Scores are calculated based on questions answered correctly for the passages read, but because the test is adaptive, these scores do not necessarily correspond to grade-level expectations for performance. Because Cluster Area Scores are based on small numbers of items per cluster, they are expressed simply using ranges: High, Medium, or Low.

Within the Broad Diagnostic Inventory in Grades K-2, Listening Comprehension and Reading Comprehension produce percent correct scores for implicit and explicit questions. In addition, Reading Comprehension produces Fluency and Words Correct per Minute scores, which are each based on passages selected according to each student's instructional reading level. The Vocabulary and Spelling tests produce only percentile rank scores.

Within the Targeted Diagnostic Inventory in Grades 3-12, Percentile Ranks and Standard Scores are available for each assessment, Maze and Word Analysis. In addition, each has a score type unique to its identity as an adaptive (Word Analysis) or non-adaptive (Maze) test. Word Analysis provides Developmental Ability Scores, while Adjusted Maze Scores, adjusted for difficulty of passages within the grade level, are available for the Maze task.

Development

The Technical Manual offers a concise and heavily referenced discussion of the foundational principles upon which the assessments were developed. “In an alphabetic orthography such as English, children must learn the conventional and intentional relations between letters and sounds—the alphabetic principle. Sensitivity to the sound structure of the spoken language to which the letters relate (i.e., phonological awareness) is an important predictor of success in mastering the alphabetic principle” (FAIR K-2 Technical Manual, p. 1). Well-documented major research findings concerning the prediction of success in learning to read were used to guide test and item development. In the early grades specifically, phonological awareness, word reading, listening comprehension, and reading comprehension predict success. Details regarding the development of specific tasks comprising the FAIR K-2 (i.e., Reading Comprehension, Vocabulary, and Spelling) are presented in the Technical Manual. A preliminary version of the FAIR was tried out in 2008-2009 using about 150 teachers and about 2,000 students. The final version incorporated changes suggested by teachers and administrators for statewide implementation in 2009-2010.

For Grades 3-12, the FAIR was developed in response to the recognized need for formative assessment throughout the academic year, the results of which can be used to identify instructional needs, monitor progress, and assist instructional efforts that will increase reading ability toward grade-level standards. The FAIR Grades 3-12 Technical Manual references an important document recently published by the National Center on Instruction concerning the importance of literacy assessment with adolescents (Torgesen & Miller, 2009) and notes that development of tasks for Grades 3-12 were responsive to two of the three purposes of literacy assessment with adolescents articulated therein. Details regarding the development of specific tasks comprising the FAIR 3-12 (i.e., Reading Comprehension, Maze Task, and Word Analysis) are presented in the Technical Manual.

Performance on FAIR items was analyzed using Item Response Theory (IRT) to determine item difficulty and item discrimination. Items were dropped from the active item pool if item difficulty was < -4.0 or $> +4.0$ or if item discrimination was < 1.0 . According to the test developers, 24 of the 461 items originally developed for the Grades 3-12 item pool were dropped. We assume that the vast majority of 24 items that were dropped had item discrimination values that were < 1.0 , because item difficulty (θ) levels tend to range between -3.0 and $+3.0$. Item retention decisions for Reading Comprehension required modification of the foregoing procedures, as passages rather than individual questions associated with the passages were evaluated. The modified procedure considered inter-rater reliability to evaluate “correct” responses and test information functions to assess information provided by each passage.

In addition, retained items were evaluated using a differential item functioning (DIF) analysis that compared performance of Black/White students, Latino/White students, Black/Latino students, students eligible/ineligible for Free or Reduced Lunch, and English Language Learners/non-English Language Learners. DIF testing was performed by grade, using the Mantel-Haenszel procedure and controlling for multiple tests with a Linear Step-Up procedure. Results are reported for Kindergarten and Grade 1 in the FAIR K-2 Technical Manual. Relatively few items demonstrated DIF for the majority of comparisons. Letter Names, Letter Sounds, and Phonological Awareness tasks had no more than 2 items flag across all comparisons and assessment periods. Word Reading had the greatest number of items showing DIF, most notably for English Language Learners as might be expected. Further, comparisons involving Latino/White (K), Black/Latino (K), and students eligible/ineligible for Free or Reduced Lunch (Grade 1) had 10, 10, and 6 items, respectively, showing DIF. Although details are not provided, the FAIR K-2 Technical Manual (p. 13) states that “Items demonstrating DIF were flagged for further study in order to ascertain why groups with the same latent ability performed differently on the items.” There is no indication of whether or how many items ultimately were removed or modified. The FAIR Grades 3-12 Technical Manual (p. 24) reports the use of similar DIF procedures, but does not present the results.

Technical

Standardization

In 2007-2008, an initial sample of students was used for item calibration and linking, where needed. This calibration sample was drawn from students in three counties: Leon, Alachua, and Manatee. Demographic information for this sample is

provided for all grades in the respective Technical Manuals. In 2008-2009, a second sample was used to gather additional data and to test the implementation of the assessment system. This implementation sample was drawn from students in the same counties as the calibration sample, plus the addition of Broward County. Demographic information for this sample is provided for Grades 3-12 in the Technical Manual but not for K-2.

For K-2, from the materials and information provided to Buros, it is not entirely clear which sample (i.e., the calibration sample or implementation sample) was used for the calculation of certain scores and norms. For Vocabulary, the test developers informed us that they “opted to create a look-up table that reflects scoring from a theoretical normal distribution of data” (B. Foorman, personal communication, April 16, 2010) to be used on an interim basis, while data are collected during the current school year that will allow empirically derived estimates to be developed. We assume that data collected during the 2009-2010 administration may be used to enhance or adjust other norms as well, given the expectation of very large numbers of students completing the FAIR over the course of 2009-2010. For Grades 3-12, the calibration sample provided the norms for the Maze task and Word Analysis, while the implementation sample was used to develop norms for the Broad Screen.

Materials provided on the Florida Center for Reading Research website (http://www.fcrr.net/files/3-12_Score_Types.pdf) describe the percentile rank norms as based upon a representative sample of Florida students. The Technical Manuals describe the samples as “roughly representative” of the Florida student population. The latter characterization may be accurate, although we noted a number of fairly large

discrepancies when grade-by-grade sample demographic data were compared to the state figures. Most concerning, a pattern of overrepresentation of Black students and underrepresentation of Hispanic students appears to characterize both the calibration and implementation samples. In each case, all but 2 of 21 comparisons of percentages followed the pattern we identified. We did not determine whether these differences differed in a statistically significant manner. In some cases, the departures were substantial (e.g., Black students comprise 23.7% of the Grade 3 population of Florida, whereas Black students comprised 54.3% and 37.0% of the calibration and implementation samples, respectively).

Reliability

Broad Screen

IRT makes it possible to report the precision of individual score points, and is particularly useful in evaluating the precision of scores at or near the cut-point. The FAIR K-2 Technical Manual provides IRT precision estimates at the cut-points for certain Broad Screen tasks, using a scale similar to that used for alpha coefficients. Reported estimates are consistently above .85, as all estimates are .86 or .87, with the exception of the first assessment period during Kindergarten. Thus, decisions based on the test scores can be rendered with confidence. We are not entirely sure how these cut-points were determined for individual tests in Kindergarten because for each assessment period in Kindergarten, two tasks are combined to produce a PRS. According to the test developers, precision estimates for the Broad Screen in Grade 2 are not appropriate to report because the test is timed.

Standard error plots across the ability scale are provided for some of the Broad Screen Tasks in the appendix of the FAIR K-2 Technical Manual. These plots are useful for examining the precision of estimates across the range of student abilities. However, because the standard errors are plotted against the IRT ability estimate rather than reported scores, we do not know how these standard errors correspond to locations of scores on the reported score metric.

Test-retest reliability estimates for the K-2 Broad Screen across assessment periods range from .44 (Kindergarten Assessment Period 1 with Assessment Period 3) to .84 (Grade 2 Assessment Period 2 with Assessment Period 3), with a median value of .69. Because each assessment period spans 34 Instructional Days and assessment periods are separated by some number of Instructional Days, the testing intervals could range from a minimum of 26 Instructional Days (between the end of Assessment Period 1 and the beginning of Assessment Period 2) to a maximum of 164 Instructional Days (between the beginning of Assessment Period 1 and the end of Assessment Period 3). The pattern of observed coefficients is as expected, with greater consistency noted at older ages and for contiguous assessment periods. Estimates would be expected to be lower as a function of the extent to which different tasks are used across assessment periods, as is especially the case in Kindergarten.

Internal consistency estimates for Grades 3-12 are expressed as generic reliability coefficients, because of the adaptive nature of the FAIR Grades 3-12. The coefficients across grades are very high, all hovering closely around .90. The test developers applied this (conservative) standard and found that only 8% of students had precision estimates

lower than .80, while 40% of students had precision estimates between .80 and .89. The remaining 52% demonstrated ability scores that met or exceeded the .90 threshold.

Broad Diagnostic Inventory (Grades K-2)

Test-retest reliability results are reported for tests in the Broad Diagnostic Inventory for Grades K-2. Across assessment periods, the vast majority of coefficients were in the moderate range – in the .40s, .50s, or .60s, probably due in part to the fact that the construct being assessed is developmental in nature, such that changes (i.e., increases) in abilities/scores would be expected as time goes by. As well, the intervals between testing sessions are not specified, but could be quite variable, given that intervals are reported by assessment periods, which themselves are quite broad.

For Listening Comprehension in Kindergarten and Reading Comprehension in Grades 1 and 2, the same passages were correlated across the three assessment periods. For Listening Comprehension (in Kindergarten), across assessment periods and passages, coefficients ranged from .49 to .63, with a median value of .565. For Reading Comprehension in Grade 1, across assessment periods and passages, coefficients ranged from .40 to .62, with a median value of .47. For Reading Comprehension in Grade 2, across assessment periods and passages, coefficients ranged from .25 to .62, with a median value of .46. Overall, test-retest correlations were slightly higher for Listening Comprehension than for Reading Comprehension. For Spelling in Grade 2, a reliability study was conducted that used alternate forms across assessment periods. The resultant coefficient was .83, quite high given the context.

Standard error plots are also provided for Comprehension and Vocabulary. The plots for Vocabulary are provided only for Assessment Period 3. For nearly the entire

range of ability and for each grade, standard errors correspond to coefficient alpha values of .80 or higher. To the extent that these standard errors are similar for Assessment Periods 1 and 2, the test can be said to produce reliable scores for assessing expressive vocabulary. Spelling standard errors are also relatively low (i.e., corresponding to a coefficient alpha of at least .80) for the range of students between two standard deviations below the mean and two standard deviations above the mean, encompassing 95% of students.

Targeted Diagnostic Inventory

Internal consistency estimates for the K-2 Targeted Diagnostic Inventory tasks are reported as alpha coefficients. In Kindergarten, coefficients range from .27 to .92, with a median value of .75. In Grade 1, coefficients ranged from $-.12^1$ to .84, with a median value of .77. In Grade 2, coefficients were .68 and .73 for Word Building and Multisyllabic Word Reading, respectively. Tasks with reliability values above .70 are generally regarded as adequate in social science contexts. Given the intended uses of the FAIR as a formative assessment used to guide the development of instructional strategies, somewhat lower coefficients are probably tolerable. Of the 18 coefficients reported for the K-2 Targeted Diagnostic Inventory tasks, 13 are greater than .65. The smallest coefficients reported are associated with the smallest sample sizes. The negative reliability coefficient observed for one of the Word Building tasks may have resulted from a small sample size ($n = 42$), coupled with the small number of items on the test ($n = 5$), both of which can lead to erratic values for reliability coefficients.

¹ Negative alpha coefficients are normally reset to 0.00 because the coefficient theoretically cannot be less than 0.

For the Word Analysis task in the Grades 3-12 Targeted Diagnostic Inventory, generic reliability was calculated using the same formula as for the Broad Screen. Coefficients exceeded .90 at each grade level, ranging from .92 to .95. These ranges are more than adequate and are aided by the adaptive administration of the test. For the Maze task in the Grades 3-12 Targeted Diagnostic Inventory, internal consistency was calculated using the split-half technique in which scores from one passage were correlated with scores from the second passage, and then corrected for length using a standard statistical (Spearman-Brown) formula for correction. For all grades, corrected coefficients were at least .80, with the exception of Grade 7 where the coefficient was .77. These values are certainly appropriate to support the intended uses of the FAIR.

Validity

Evidence supporting the validity of test scores must be evaluated in tandem with the intended uses of the test scores. The stated purposes of the FAIR system center on evaluating progress towards the end-of-year benchmarks, diagnosing learning needs and setting instructional goals accordingly, and monitoring instructional progress. Simply put, “the purpose of these tests is to provide data at the beginning and throughout the school year for use by teachers, schools, and district leaders to strengthen instruction for all students, but particularly for students who may struggle to meet grade level standards in reading” (FAIR Grades 3-12 Technical Manual, p. 1).

For tests of educational achievement, a major source of validity evidence is based on test content. A second source of evidence derives from relationships to other variables that may be demonstrated through correlations between the FAIR tests and other widely-recognized measures of the same skills. A third source that also derives from

relationships to other variables includes a criterion measure that FAIR test scores may predict. In this case, studies that examine the relationship between FAIR test scores and end-of-year tests (i.e., SESAT and SAT-10 for K-2, and the FCAT Reading test for Grades 3-12) were used to provide evidence of the predictive validity of test scores.

Grades K-2

In Kindergarten, the Broad Screen is used to predict performance on the end of year SESAT test, and in Grades 1 and 2, the Broad Screen is used to predict performance on the end of year SAT-10. As previously described, the score associated with the Broad Screen in Grades K-2 is the Probability of Reading Success (PRS), which quantifies students' likelihood of scoring at or above the 40th percentile on the SESAT or SAT-10. The FAIR K-2 Technical Manual presents a series of 2 x 2 contingency tables that compare students' classification on the Broad Screen (i.e., above or below a PRS score of 85%) to whether the student later scored above or below the 40th percentile on the end-of-year test. The publishers chose to use negative predictive power to evaluate the strength of the predictive relationship. Negative predictive power reflects the percent of students correctly identified as not-at-risk by the screening measure, using the end-of-year assessment results as the comparison. The test developers established a target goal of 85% negative predictive power, meaning that 85% of students classified according to their PRS or FSP scores as not-at-risk would end up not-at-risk on the outcome measure (SESAT, SAT-10, or FCAT). The test met the established criterion outright in Assessment Period 1. Negative predictive power for Assessment Periods 2 and 3 were .81 and .79, respectively. In Grade 1, negative predictive power approached or surpassed .90 at each assessment period. In Grade 2, negative predictive power met the criterion in

Assessment Period 3 (.86), and nearly met the criterion in Assessment Period 2 (.84) and Assessment Period 1 (.81).

The Technical Manual describes a second validity study relating to prediction, in which PRS scores were entered as a single predictor in a logistic regression analysis. Logistic regression is used to calculate, for given values of a predictor or of multiple predictors, the probability of a yes-or-no outcome; as well, it can be used to calculate the percent of variance in the outcome accounted for by the predictor or predictors. The outcome variable in these analyses was the end-of-year test, SESAT Word Reading subtest for Kindergarten and SAT-10 Reading Comprehension Subtest in Grades 1 and 2. For Kindergarten, the percent of variance in the outcome that is accounted for by the predictor(s) in the model was 17%, 26%, and 23% for Assessment Periods 1, 2, and 3, respectively. For Grade 1, percent of variance accounted for was 41%, 46%, and 56% for the respective Assessment Periods. For Grade 2, percent of variance accounted for was 23%, 26%, and 29% for the respective Assessment Periods. For a short test, these are sizeable percent values. Sample sizes are not specified for these analyses; we presume they are the same samples shown for the 2 x 2 classification contingency tables.

Although certain groups may be classified as at-risk more than other groups, an important piece of information for evaluating test bias is whether the test predicts performance as well for members of these groups as for students who are not members of these groups. The FAIR K-2 Technical Manual presents a series of 12 tables summarizing findings from analyses that examined differential accuracy of the regression equations across different demographic groups. For each grade, results are presented for Black students, Latino students, students who are English Language Learners (ELL), and students eligible for Free/Reduced Lunch. Sample sizes ranged from 230 to 506. A

significant interaction would indicate that the accuracy of prediction was significantly different for members of these groups than for those not part of the groups.

Results for Kindergarten and Grade 2 indicate no significant interactions between risk status as determined by the Broad Screen and any of the four demographic characteristics. In Grade 1, however, significant interactions were noted between risk status and Latino/non-Latino students at Assessment Period 1, between risk status and ELL at Assessment Period 1, and between risk status and Free/Reduced Lunch eligibility at Assessment Period 3. In addition, the interaction between risk status and Latino/non-Latino student approached significance ($p = .06$) at Assessment Period 3. These results raise concerns about the utility of the FAIR in Grade 1, especially for Latino students who also may be learning English. The test developers acknowledge the need to examine these results more closely as they relate to differential validity (FAIR Grades K-2 Technical Manual, p. 32). Otherwise, these results support the use of the Broad Screen with members of diverse populations in Grades K-2, within the caveats noted in an earlier section of this review regarding results of DIF analyses.

Concurrent evidence of test score validity derives from a study in which a sample of students ($n = 370$) from Grades K, 1, and 2 completed the FAIR and the Expressive Vocabulary Test, 2nd edition (EVT-2; Williams, 2007). Correlations between the EVT-2 and the FAIR Vocabulary test were moderately high, ranging from .75 to .83 across grades.

Grades 3-12

For Grades 3-12, the Broad Screen is used to predict performance on the FCAT Reading test. As previously described, scores on the Broad Screen at each assessment period are combined with performance on the prior year's FCAT score to produce FCAT

Success Probability (FSP) scores, which quantify students' likelihood of scoring at or above Level 3 on the FCAT Reading test administered at year's end. (The exception is Grade 3, which does not have a previous year's FCAT score.) Procedures for examining the validity of the Broad Screen follow those of Grades K-2.

The FAIR Grades 3-12 Technical Manual presents a series of 2 x 2 contingency tables that compare students' risk classification on the Broad Screen (i.e., FSP > 85% and FSP < 85% for not-at-risk and at-risk, respectively) to whether the student later scored at or above grade level on the end-of-year test (i.e., Level 3 or higher on FCAT Reading). The publishers chose to use negative predictive power to evaluate the strength of the predictive relationship, noting that a goal of 85% negative predictive power was sought. Negative predictive power ranged from .54 (Grade 10) to .98 (Grades 3, 4, and 5), with a median value of .925. Across grade levels, the test met the established criterion as negative predictive power was at least .85, with the exception of Grade 10, where the sample of students classified as not-at-risk was very small ($n = 13$). Consequently, for most grades, there is only a small chance that a student classified as not-at-risk based on his or her FSP score will fail to reach Level 3 or above on the current year's FCAT.

The FAIR Grades 3-12 Technical Manual also presents a table that lists correlations between (a) Broad Screen performance and current year FCAT score and (b) prior year FCAT score and current year FCAT score. Coefficients were relatively high, ranging from .62 to .74 and from .75 to .83, respectively. We requested and received more information from the Florida Department of Education about the combination of the Broad Screen and prior year's FCAT in predicting performance on the current year's FCAT. For each grade, once prior year's FCAT is included in the logistic regression,

adding Broad Screen improves prediction only slightly. However, one benefit of including the Broad Screen in the FSP score is that students can show improvement on the Broad Screen throughout the year, which of course is not possible with the previous year's FCAT score. In fact, predicting test performance is somewhat incidental to the stated purposes of FAIR, which center on instructional improvement for all students, but especially for those struggling to meet grade level expectations in reading.

We also requested but did not receive the specific weights used to calculate the FSP score from its two component variables. Apparently, this information is deemed proprietary (B. Foorman, personal communication, May 3, 2010). Therefore, we cannot comment specifically upon how much the Broad Screen measure itself contributes to the prediction of end-of-year success on the FCAT Reading test.

Regarding differential accuracy of prediction, the FAIR Grades 3-12 Technical Manual presents the results of a series of logistic regressions predicting FCAT Reading success, based on risk status as determined by FSP scores. Demographic variables examined included Black and Latino students, students identified as English Language Learners, and students eligible for Free/Reduced Lunch. No significant interactions were observed between risk status and the demographic characteristics. Thus, the evidence provided supports the utility of the FAIR Broad Screen with members of various diverse populations in Grades 3-12.

In addition, the test developers offer evidence of score validity for the Maze and Word Analysis tasks, by demonstrating that students who perform at different FCAT levels also perform differently on these FAIR tasks.

Commentary

FAIR is a very recently published test (2009), developed for a highly specific target population and purpose. It appears to have been developed with careful attention to many important test development standards. Overall, we commend the FAIR test developers for their efforts and accomplishments in creating the FAIR.

In the context of a newly developed measure, the documentation accompanying the FAIR is of exceptionally high quality, and we expect that a few mostly minor errors or omissions we noticed will be corrected as the test, its manuals, and other materials are further refined and developed. For example, it would be desirable to include sample sizes along with tabulated information throughout the Technical Manuals. Also, it would be desirable to provide more specific information about the development of percentile rank norms for tasks that ordinarily are not taken by all students at each assessment period – Vocabulary and Spelling in K-2 and the Maze task and Word Analysis in Grades 3-12.

FAIR was developed specifically for use within the state of Florida. We regard this fact as beneficial to Florida's students. The test developers have a vested interest in the quality of education within the state and also have firsthand knowledge of Florida's standards as reflected in the FCAT. They appear invested in assessment practices that go beyond merely responding to accountability demands – although accountability is certainly an essential outcome. By design, the FAIR provides formative information and offers a variety of tasks and tools intended for use between assessment periods – that is, for instructional purposes. The development of ongoing progress monitoring tasks speaks directly to this issue. These latter uses respond to the needs of individuals in the classroom – students and teachers.

We applaud the combination of FAIR-derived and previous year's FCAT scores toward the end of identifying students who are at risk. One of the most commonly cited maxims in testing is that no single score should be used alone by itself. Nevertheless, under No Child Left Behind, decisions are often made on this basis. Therefore, we believe that the approach embodied in the FAIR is a well informed one.

The inclusion of Lexile Measures and Ranges should prove useful in assisting teachers and librarians in selecting texts for students that are appropriate for their reading levels. We are impressed that the FAIR test developers undertook the requisite linking study. Because the Broad Screen is administered at three assessment periods throughout the academic year, student progress reflected in this scale can be used to help teachers and librarians to continually adjust text selections accordingly.

It is our understanding and expectation that during the first year or two of use (and possibly longer), data collected through the PMRN data entry system will be used to refine and update technical information for the FAIR. This approach would be a responsible course of action as doing so could strengthen an already good test by increasing both the size and representativeness of its reference groups.

In considering future development work on the FAIR, we note that evidence to support the use of "approved accommodations" is most certainly needed for both the FAIR K-2 and the FAIR Grades 3-12. It appears that the test developers relied upon a subjective approach to determine which accommodations they thought would not affect the interpretations or utility of resultant scores. Ultimately, however, the impact of accommodations is an empirical question the answer to which requires the accumulation

of data showing how score interpretations remain valid after given accommodations have been implemented.

Examination of the standard error plots presented in the appendix of the FAIR K-2 Technical Manual leads us to agree with the caveat presented in the Technical Manual (p. 16) that the Comprehension tasks provide “descriptive information about a student’s emerging comprehension skills rather than reliable quantitative information.” We were pleased to confirm that this limitation is mentioned in the training materials for teachers and other individuals who work with FAIR scores (B. Foorman, personal communication, May 3, 2010).

Negative predictive power was used consistently to evaluate the accuracy of classification decisions. As cited in the Technical Manuals, other methods of evaluating classification decisions include positive predictive power, sensitivity, and specificity. We agree with the test publisher that, for FAIR, the consequences of failing to identify students as at-risk in time to provide needed instruction and intervention is the most important classification error to avoid. Whenever a test is used, there will always be false positives, as well as false negatives. Each has its own costs. For example, costs for falsely identifying students as at-risk include teachers spending extra time with students who do not need extra help and having less time to spend with students who do need additional assistance. Although not explicitly addressed in the test materials, we trust these costs were at least implicitly considered when choosing the PRS/FSP cut score of 85%. In our opinion, it would be beneficial to address costs associated with false positives as well as false negatives in the FAIR Technical Manuals.

Summary

The FAIR is a recently developed reading measure currently in use throughout the state of Florida in Grades K through 12. The primary purpose of FAIR is to provide formative assessment that can be used to guide instruction that ultimately improves reading education for all students. Scores representing reading success probabilities are derived from performance on the FAIR, and are used to gauge the likelihood that students will perform successfully on the end-of-year reading assessment associated with their grade level. Test development efforts are to be commended. In general, psychometric evidence is relatively complete and supports the test's intended uses. The test developers appear mindful of the need to continue their development efforts, a process that will be ongoing and feasible due to the corresponding development of a system wide data entry system.

References

- Florida Department of Education. (2009). *Florida Assessments for Instruction in Reading administration manual: Grades 3-12*. State of Florida, Department of Education.
- Florida Department of Education. (2009). *Florida Assessments for Instruction in Reading administration manual: Kindergarten – grade 2*. State of Florida, Department of Education.
- Florida Department of Education. (2009). *Florida Assessments for Instruction in Reading technical manual, 2009-2010 edition: Grades 3-12*. State of Florida, Department of Education.
- Florida Department of Education. (2009). *Florida Assessments for Instruction in Reading technical manual, 2009-2010 edition: Kindergarten – grade 2*. State of Florida, Department of Education.
- Torgesen, J. K., & Miller, D. H. (2009). *Assessments to guide adolescent literacy instruction*. Portsmouth, NH: RMC Research Corporation, Center on Instruction.
- Williams, K. (2007). *Expressive Vocabulary Test*, 2nd edition. San Antonio, TX: Pearson.