School Psychology Review,
2010, Volume 39, No. 3, pp. 463–483

An Empirical Review of Psychometric Evidence for the Dynamic Indicators of Basic Early Literacy Skills

Catherine T. Goffreda and James Clyde DiPerna
The Pennsylvania State University

Abstract. The Dynamic Indicators of Basic Early Literacy Skills (DIBELS) are brief measures of early literacy skills for students in Grades K–6 (University of Oregon, 2009; see Kaminski & Good, 1996). School psychologists and other educational professionals use DIBELS to identify students who are in need of early intervention. The purpose of this review was to synthesize the current psychometric evidence for each DIBELS indicator. Strong reliability and validity evidence was observed for DIBELS Oral Reading Fluency; however, evidence for the remaining DIBELS indicators demonstrated greater variability. Although the majority of evidence focused on individual score reliability and validity for single-point decisions, further studies are needed to determine effective practices for progress monitoring.

In the current era of educational accountability, teachers and administrators are increasingly proactive in identifying and providing interventions for students at risk for reading failure (Good, Kaminski, Smith, Simmons, Kame'enui, & Wallin, 2003). Extensive evidence indicates that without early intervention, students with deficits in early literacy skills experience poor learning trajectories of reading growth (e.g., Cunningham & Stanovich, 1998; Juel, 1988; Stanovich, 1986). Thus, efficient measures are essential to monitor young children's development of early literacy skills (Fuchs & Fuchs, 1999). In response to this need, progress monitoring methods such as curriculum-based measurement (CBM) have been developed to follow students' academic skill development and identify students who may benefit from early intervention.

Alphabetic knowledge, phonemic awareness, and fluency are strong predictors of future literacy performance (Adams, 1990; Good, Simmons, & Kame'enui, 2001; Lundberg, Frost, & Petersen, 1988; Torgesen, 2002). Alphabetic knowledge involves associating letters with corresponding sounds (Byrne & Fielding-Barnsley, 1989). Phonemic awareness incorporates the skills of isolating, blending, and segmenting words into phonemes, or the ability to manipulate individual sounds within a word (Blachman, 1991). Fluency refers to automaticity, or reading at an appropriate pace with little cognitive effort (Hasbrouck, 1998). As such, these factors should be incorporated into early literacy progress monitoring measures to maximize predictive validity.

One of the most frequently used measures for screening and/or progress monitoring of these early literacy skills is the Dynamic

Correspondence regarding this article should be addressed to Catherine T. Goffreda and James Clyde DiPerna, College of Education, The Pennsylvania State University, 226 CEDAR Building, University Park, PA 16802; E-mail: ctg130@psu.edu or jcd12@psu.edu

Copyright 2010 by the National Association of School Psychologists, ISSN 0279-6015

Indicators of Basic Early Literacy Skills (DIBELS; University of Oregon, 2009; see Kaminski & Good, 1996). DIBELS encompasses a set of brief standardized measures for students in the primary grades (K–3), with additional measures available for students in the intermediate grades (4–6). DIBELS was developed to assess three of the key early literacy domains (phonological awareness, alphabetic understanding, and fluency) identified by the National Reading Panel (2000). Scores from DIBELS have been linked to reading fluency in later elementary years, which enables educators to identify and provide early intervention to struggling students at risk for future problems (Kaminski & Good, 1996). Like CBM, DIBELS is a form of general outcome measure, or a brief standardized measure with parallel forms that can be used to assess global skill growth over time (Hintze, Christ, & Methe, 2006). However, DIBELS differs from reading CBM in that it features standardized content rather than using materials sampled from a district's curriculum. Kaminski and Good (1996) describe the measure as "dynamic" because prereading skills are assessed on a continual basis and as "indicators" because they measure key components of basic early literacy skills.

DIBELS benchmark assessments are administered three times per year (fall, winter, spring). Scores can be compared to empirically derived decision categories for a student's grade level, thus allowing educators to identify struggling students and provide an appropriate reading intervention. *Low-risk* performance indicates that students have an 80% chance of achieving future proficiency. *Some-risk* performance denotes that students have a 50% chance of achieving future proficiency. Finally, *at-risk* performance indicates that students have an 80% chance of *not* achieving future proficiency (University of Oregon, 2009). In addition to the benchmark assessments, DIBELS includes multiple parallel forms of each measure that can be used monitor skill progress for students receiving additional instructional support.

DIBELS Overview

DIBELS (Good & Kaminski, 2002) consists of five core indicators, each measuring a fundamental early literacy skill: Initial Sound Fluency (ISF), Letter Naming Fluency (LNF), Phoneme Segmentation Fluency (PSF), Nonsense Word Fluency (NWF), and Oral Reading Fluency (ORF). Two supplemental measures, Word Use Fluency and Retell Fluency, have been added to the DIBELS to measure vocabulary/oral language, and reading comprehension, respectively. However, given that these measures are supplemental and not included in any of the benchmark assessments, they were not included in this review.

ISF is used to measure phonemic awareness through the winter benchmark of Grade K. An examiner presents four pictures per item, and the student must identify which picture has the same initial sound as an orally presented stimulus (e.g., "Choose the word beginning with the *g* sound"). Sixteen sets of four pictures are administered at each benchmark. The student's score is the number of correct initial sounds per minute.

LNF is used to assess knowledge of the alphabetic letters in kindergarten, as well as in the fall of Grade 1. The student is presented with a page of upper- and lowercase letters, and verbally names as many as possible. The student's score is the total number of correct letters named within 1 min.

PSF is used as an applied measure of phonemic awareness skills. This indicator is administered in the winter and spring benchmarks of kindergarten, as well as all Grade 1 benchmarks. The examiner orally presents a monosyllabic word (e.g., "cat"), which the student must segment into individual phonemes (e.g., "/c/ /a/ /t/"). The total score is the total number of correct phonemes produced within 1 min.

NWF is used to measure the alphabetic principle, including alphabetic understanding and phonological recording. NWF is administered in the winter (optional) and spring benchmarks of kindergarten, all benchmarks of Grade 1, and the fall benchmark of Grade 2. The student is presented with a sheet of simple

Vowel Consonant and Consonant Vowel Consonant nonsense word sequences (e.g., "tob," "rup," "kud," etc.) and may either pronounce individual letter sounds or the entire word. Each score is the number of correct letter sounds produced within 1 min.

ORF is used to assess a child's fluency in reading connected text. Beginning in winter of Grade 1, this indicator is administered to students in both the primary (1–3) and intermediate (4–6) grades. Students read standardized grade-appropriate passages aloud to the examiner. The score, or ORF rate, is the number of words read correctly per minute.

Purpose and Rationale

Since the release of the seminal National Research Council (1998) and National Reading Panel (2000) reports, use of DIBELS has experienced exponential growth in elementary schools throughout the United States. In 2003, Brunsman reported that the online DIBELS data management system included 300 school districts, 600 schools, and 32,000 children. Just 5 years later, over 15,000 schools used the K–3 DIBELS Data System (University of Oregon, 2009). Legislation such as the No Child Left Behind Act of 2001 (PL 107–110) has placed an increasing emphasis on school accountability for academic achievement; thus, gauging academic performance to predict performance on annual district and state standardized assessments has become paramount.

Although the extant literature includes multiple empirical studies of DIBELS' reliability, validity, and classification accuracy, the evidence is distributed across a variety of publication types (e.g., technical reports, peer-reviewed outlets, dissertations/theses) and varies by indicator. In addition, previous comprehensive literature reviews regarding the technical adequacy of CBM reading measures (e.g., Wayman, Wiley, Ticha, & Espin, 2007) examined specific measures such as reading aloud, maze, and word identification, but did not examine evidence for specific "families" of general outcome measures such as DIBELS. Finally, DIBELS is not without its critics (e.g., Samuels, 2007) and standards for educational assessment by the American Educational Research Association (AERA), American Psychological Association (APA), and National Council for Measurement in Education (NCME; 1999) require that practitioners evaluate the psychometric data for the assessments used and for the populations with which the data are collected. The widespread use of DIBELS suggests that the reliability of the data and the validity of resulting screening and progress monitoring decisions should be carefully examined. Thus, the purpose of this review was to synthesize the current published psychometric evidence (reliability, validity, and classification accuracy) for each DIBELS indicator. Specifically, the following research questions guided this review:

1. How reliable are DIBELS scores for various decision-making purposes (i.e., screening, group outcomes, individual outcomes)?
2. How valid are decisions made from DIBELS data when measuring primary grade students' early literacy skills?
3. How accurate are DIBELS scores in predicting subsequent diagnostic decisions of reading proficiency?
4. To what extent do DIBELS reliability and validity coefficients vary across racial and ethnic groups?

Determining the psychometric adequacy of individual DIBELS measures for various assessment purposes (i.e., screening, group, and individual decision making) will further help inform practitioners of best practices for early identification of at-risk students. As such, our primary goal was to not only identify evidence gaps and directions for future research, but also to assist practitioners utilizing these measures in applied settings.

Method

Literature Review

A variety of strategies were employed to identify extant empirical literature regarding DIBELS' technical adequacy. First, an inclusive list of known citations referencing DIBELS or general literacy indicators was

accessed from the DIBELS website (https://dibels.uoregon.edu). This document reported 88 citations, consisting of 20 peer-reviewed articles, 8 book chapters, 39 dissertations/theses, 16 technical reports, 1 poster presentation, and 4 manuscripts submitted for publication.

Following the review of the DIBELS website, follow-up searches on PsycINFO and ERIC were conducted using the key word "DIBELS," which yielded 89 potential references. Of these, only 16 were included on the aforementioned list of DIBELS citations. The 73 remaining citations (28 peer-reviewed articles, 39 dissertations, and 6 book chapters/reports) were added to the 88 citations from the DIBELS Web site (161 total citations). Finally, searching the key words "Dynamic Indicators of Basic Early Literacy Skills" yielded 67 results. Of these, 37 citations had been identified previously. The 30 new citations (14 peer-reviewed articles and 16 dissertations) were added to the "initial review" list for a total of 191 unique citations related to DIBELS. All included citations were carefully reviewed to avoid duplication of data across dissertation/theses, technical reports, and peer-reviewed articles. Final searches were completed in December 008. Thus, this review does not include studies published since that time.

Inclusion/Exclusion Criteria

To be included in the review, studies had to report reliability, validity, and/or classification accuracy statistics for at least one of the five DIBELS early literacy measures. Based on these criteria, 165 publications were eliminated from our list of 191 unique DIBELS citations. These publications were eliminated for the following reasons:

1. Thirty-seven were excluded because the articles did not report any DIBELS data (e.g., descriptive book chapters).
2. Fifty-six publications were excluded because they did not report psychometric data (i.e., reliability coefficients or comparison to a measurable criterion such as early literacy and/or reading measures with published psychometric evidence).
3. Twenty-one publications were excluded because they used progress monitoring measures other than DIBELS (e.g., CBM-ORF).
4. Fifteen publications were excluded for using a design that involved fewer than 25 participants because they did not report the statistical indices (reliability estimates, validity coefficients) featured in this review.
5. Finally, 36 publications were inaccessible or unavailable for inclusion. We made at least two attempts to obtain dissertations, theses, and peer-reviewed articles via interlibrary loan. If these efforts were unsuccessful, we attempted to directly contact the authors via e-mail. Of these, 20 were unpublished student dissertations and theses, 7 were technical reports, 4 were manuscripts submitted for publication, 2 were book chapters, 2 were poster presentations, and 1 was a peer-reviewed journal article.

The 26 remaining references, including 13 peer-reviewed articles, 7 dissertations, and 6 technical reports, were included in the following review.

Analytic Approach

Each of the 26 studies was reviewed for quantitative reliability, validity, and decision-making accuracy data. Reliability evidence, or the dependability and consistency of scores (AERA, APA, & NCME, 1999), was classified according to three common forms (test–retest, alternate-form, and inter-rater reliability). Although various criteria are used in the literature for interpretation of reliability evidence, no standards have been universally endorsed. However, the guidelines described by Salvia, Ysseldyke, and Bolt (2010) are frequently cited in the school psychology and special education literatures. Thus, they were utilized in this study. Salvia et al. (2010) recommended reliability standards of at least .60 when scores are reported for groups, .70 for

frequent (at least weekly) progress monitoring, .80 for screening, and .90 for important individual decisions.

Validity is the extent to which a test measures what it is intended to measure (Messick, 1995). Currently, the aforementioned standards for educational assessment (AERA, APA, & NCME, 1999) specify five important sources of validity evidence, test content, response processes, internal structure, relationship to other variables, and consequences of testing. However, our searches only yielded studies that focused on the relationships between DIBELS scores and other variables. As such, we classified this evidence as either convergent or criterion-related (including concurrent and predictive) sources of validity. Specifically, only comparisons to assessments measuring a theoretically identical construct (e.g., phonemic awareness for PSF) were classified as convergent validity evidence. Comparisons to assessments measuring a theoretically similar but not identical assessment (e.g., oral reading for PSF) were classified as criterion-related validity evidence. Comparisons of DIBELS indicators to outcome measures in the future (e.g., statewide academic achievement tests) were classified as predictive validity evidence. For descriptive purposes, validity coefficients were classified based on the Wayman, Wallace, Wiley, Ticha, & Espin (2007) ranges: $r > .70$ (high); $.50 < r < .69$ (moderate); and $r < .49$ (low).

Finally, decision-making accuracy (i.e., accuracy of diagnostic decisions based on DIBELS scores) was also reviewed as a form of predictive validity evidence. Specifically, four types of evidence were considered: sensitivity, specificity, positive predictive power, and negative predictive power. Swets (1988) recommended that sensitivity and specificity levels exceeding 75% should be considered adequate. In the context of the current review, sensitivity denotes the percentage of students *not* proficient on an outcome measure of reading skill who were identified as *at risk* by DIBELS indicators. Specificity denotes the percentage of students proficient on an outcome measure who were identified as *low risk* by the DIBELS. Conversely, positive predictive power refers to the percentage of students identified as *at risk* based on DIBELS who did *not* meet proficiency on an outcome measure. Negative predictive power refers to the percentage of students identified as *low risk* by a DIBELS indicator who did, in fact, meet proficiency on a reading outcome measure (Hambleton, Swaminathan, Algina, & Coulson, 1978; Livingston & Lewis, 1995).

Although screening (i.e., single-point decision) and progress monitoring (i.e., evaluation of growth over time) purposes are not mutually exclusive, the nature of these assessment practices requires different sources of psychometric evidence. Fuchs (2004) suggested a three-stage framework for conceptualizing research evidence necessary to justify use of scores from progress monitoring measures. First, *Stage 1* involves technical features of the static score (i.e., at a single point in time). *Stage 2* refers to technical features of the slope, and *Stage 3* encompasses instructional utility (i.e., academic-related decisions and student achievement). Fuchs further emphasized that, although CBM research has focused disproportionately on Stage 1, it is "an important first step in validating a CBM system" (p. 191).

Results

To identify the relative strengths and weaknesses of psychometric evidence for each measure, the results section has been organized by DIBELS indicator.

ISF

Reliability and validity evidence. Psychometric evidence for ISF is reported in Table 1. Only one study included in the current review reported alternate-form reliability for ISF, and no studies reported test–retest or inter-rater reliability coefficients. Similarly, published validity evidence for ISF was limited to two studies. Both studies reported concurrent validity, with coefficients in the low range. Although based on a single study, convergent validity was moderate. No reviewed studies, however, reported predictive validity coefficients for ISF.

Table 1
Reliability, Validity, and Decision-Making Evidence for ISF

		Reliability Evidence			Validity Evidence			Decision Making (%)			
Authors	Sample	Test–Retest	Alt. Form	Inter-Rater	Convergent	Concurrent	Predictive	Sens.	Spec.	Pos.	Neg.
Clarke et al. (2003)	Grades K–1 (N = 27) 89% AA, 11% O ≥80% FCL					−.03[a]*					
Hintze, Ryan, & Stoner (2003)	Grade K (N = 86) 93% C, 2% AA, 2% Hispanic, 3% O 39% FCL		.86		.60[b]	.46[c] .20[d]		100[b] 91[c] 90[d]	39[b] 36[c] 37[d]	26[b] 17[c] 17[d]	100[b] 96[c] 96[d]

Note. Studies were peer reviewed. ISF = Initial Sound Fluency; Alt. = alternate; Sens. = sensitivity; Spec. = specificity; Pos. = positive predictive power; Neg. = negative predictive power; AA = African American; O = Other; C = Caucasian; CTOPP = Comprehensive Test of Phonological Processing; FCL = eligible for a free or reduced-cost lunch.
[a] Kindergarten Reading Engagement Scale.
[b] CTOPP Phonological Awareness Composite.
[c] CTOPP Phonological Memory Composite.
[d] CTOPP Rapid Naming Composite.
* Correlation was nonsignificant. Correlations were significant unless otherwise noted.

Accuracy and diversity considerations. Evidence for ISF decision-making accuracy is currently limited to one study, and results indicated adequate levels of sensitivity and negative predictive power. Specificity and positive predictive power, however, were substantially lower. Only one study (Clarke, Power, Blom-Hoffman, Dwyer, Kelleher, & Novak, 2003) reported data for students of minority status. This study included African American students, and concurrent validity was slightly lower than for studies featuring samples of primarily Caucasian students.

LNF

Reliability and validity evidence. Psychometric evidence for LNF is reported in Table 2. Four peer-reviewed studies reported reliability evidence for LNF probes. Of these, test–retest reliability coefficients met criteria for screening purposes, and in some cases, group decision making, across all studies. Alternate-form reliability, reported in two studies, was adequate for screening as well. Only two studies reported inter-rater reliability, which similarly met criteria for both screening and individual decision-making purposes (Salvia et al., 2010). Reliability evidence (i.e., test–retest, alternate form, and inter-rater) for LNF was fairly robust across multiple indices, indicating that the probes are consistent measures of student performance across time periods, forms, and examiners.

Although none of the studies reported convergent validity evidence for LNF, concurrent validity coefficients (featuring 21 different literacy measures) ranged widely from low to high ($Mdn = .54$) across the five peer-reviewed articles. Only two studies reported predictive validity evidence, with coefficients ranging from low to moderate.

Accuracy and diversity considerations. Two peer-reviewed articles reported decision-making accuracy statistics for LNF using a variety of reading outcome measures. As with ISF, negative predictive power was adequate but positive predictive power was fairly low. Sensitivity and specificity varied widely depending on the criterion measure. Two studies (Clarke et al., 2003; Riedel, 2007) reported data for students of minority status. Both studies included African American students, and coefficients did not differ from studies consisting of primarily Caucasian samples.

PSF

Reliability and validity evidence. Psychometric evidence for PSF is reported in Table 3. Two peer-reviewed articles and one dissertation reported reliability evidence for PSF. Of these, one study reported test–retest reliability coefficients, which exceeded standards for screening and group decision-making purposes. Two studies reported alternate-form reliability coefficients that were adequate for screening to individual decision-making purposes. Inter-rater reliability evidence for PSF was not reported in any studies included in the review.

Convergent validity evidence was reported in only one study, which yielded a moderate relationship. However, five peer-reviewed articles and two dissertations reported concurrent validity evidence with 22 different literacy measures. Coefficients across studies ranged from low to high ($Mdn = .33$). Predictive validity coefficients, reported in three articles and one dissertation, ranged from low to moderate ($Mdn = .38$).

Accuracy and diversity considerations. Similar to the evidence for LNF, only two peer-reviewed articles examined decision-making accuracy for PSF. As indicated in the tables, sensitivity and negative predictive power were adequate for PSF, whereas specificity and positive predictive power were significantly lower. As with LNF, Riedel (2007) was the only study that included data for students of minority status. Coefficients did not differ from studies consisting of primarily Caucasian samples.

NWF

Reliability and validity evidence. Psychometric evidence for NWF is reported in Table 4. Two studies reported test–retest co-

Table 2
Reliability, Validity, and Decision-Making Evidence for LNF

		Reliability Evidence			Validity Evidence			Decision Making (%)			
Authors	Sample	Test–Retest	Alt. Form	Inter-Rater	Convergent	Concurrent	Predictive	Sens.	Spec.	Pos.	Neg.
Clarke et al. (2003)	Grades K–1 (N = 27) 89% AA, 11% O ≥80% FCL					.75[a]					
Elliott, Lee, & Tollefson (2001)	Grade K (N = 75) 63% C, 37% O 36% FCL	.90	.80	.94		.63–.75[b] .12–.54[c] .63[d] .67[e] .50[f]					
Hintze et al. (2003)	Grade K (N = 86) 93% *C*, 2% *AA*, 2% H, 3% O 39% FCL		.94			.53[g] .52[b] .58[i]		87[g] 100[h] 100[i]	34[g] 35[h] 36[i]	22[g] 18[h] 18[i]	97[g] 100[h] 100[i]
Jordan, Kaplan, Oláh, & Locuniak (2006)	Grade K (N = 411) 43% C, 36% AA, 16% H, 5% O	.93									
Kaminski & Good (1996)	Grades K–1 (N = 78) 99% C, 1% O 12% FCL	.93[k] .83 1st				.13–.59[j] .77[k] .27–.67[l] .45[m] .34–.85[n] .50[o]					
Kamps et al. (2003)	Grades K–2 (N = 383) 40% AA, 34% C, 8% H, 18% O					.74[q] .79[p]					

(Table 2 continues)

Table 2 continued

Authors	Sample	Reliability Evidence			Validity Evidence			Decision Making (%)			
		Test–Retest	Alt. Form	Inter-Rater	Convergent	Concurrent	Predictive	Sens.	Spec.	Pos.	Neg.
Riedel (2007)	Grade 1 (N = 1,518)					.15[1r]		68[r]	65[r]		
	92% AA, 8% O					.44[2r]		67[s]	64[s]		
	85% FCL; 4% ELL					.40[s]					
Schilling et al. (2007)	Grade 1 (N = 2,588)						.30–.57[t]				
	60% AA, 25% C,										
	13% H, 2% O										
	81% FCL										
Greene (2002)	Grade K (N = 25)						.32–.39[u]				

Note. LNF = Letter-Naming Fluency; Alt. = alternate; Sens. = sensitivity; Spec. = specificity; Pos = positive predictive power; Neg. = negative predictive power; AA = African American; O = Other; H = Hispanic; C = Caucasian; FCL = Eligible for a free or reduced-cost lunch. ELL = English language learners; CTOPP = Comprehensive Test of Phonological Processing; DIBELS = Dynamic Indicators of Basic Early Literacy Skills.

[a] Kindergarten Reading Engagement Scale.
[b] Woodcock-Johnson Psychoeducational Achievement Battery—Revised.
[c] Kaufmann Brief Intelligence Test.
[d] Teacher Rating Questionnaire (Prereading).
[e] Developing Skills Checklist.
[f] Test of Phonological Awareness.
[g] Comprehensive Test of Phonological Processing (CTOPP) Phonological Awareness Composite.
[h] CTOPP Phonological Memory Composite.
[i] CTOPP Rapid Naming Composite.
[j] McCarthy Scales of Children's Abilities.
[k] Metropolitan Readiness Test.
[l] Rhode Island Pupil Identification Scale.
[m] Curriculum-based measurement.
[n] Teacher Rating Scale.
[o] Stanford Diagnostic Reading Test.
[p] DIBELS Nonsense Word Fluency.
[q] DIBELS Oral Reading Fluency.
[1r] Group Reading Assessment and Diagnostic Evaluation (English-Speaking Students).
[2r] Group Reading Assessment and Diagnostic Evaluation (ELL Students).
[s] TerraNova CAT Reading.
[t] Iowa Test of Basic Skills.
[u] Early Screening Profile.

Table 3
Reliability, Validity, and Decision-Making Evidence for PSF

Authors	Sample	Reliability Evidence			Validity Evidence			Decision Making (%)			
		Test–Retest	Alt. Form	Inter-Rater	Convergent	Concurrent	Predictive	Sens.	Spec.	Pos.	Neg.
					Peer Reviewed						
Burke & Hagan-Burke (2007)	Grade 1 (N = 213) 53% C, 24% AA, 6% H, 17% O 33% FCL					.26–.33[a]					
Good et al. (2001)	Grades K–3 (N = 302–378) 90% C, 10% O 37–63% FCL						.38[b]	89[c]	92[c]		
Hintze et al., (2003)	Grade K (N = 86) 93% C, 2% AA, 2% H, 3% O 39% FCL		.97		.53[d]	.39[e] .09[f]		93[d] 100[e] 80[f]	23[d] 23[e] 21[f]	20[d] 16[e] 13[f]	96[d] 100[e] 88[f]
Jordan et al. (2006)	Grade K (N = 411) 43% C, 36% AA, 15% H, 6% O	.88									
Kamii & Manning (2005)	Grades K–1 (N = 208) 85–95% C, 4–10% AA, 1–5%, O							.59[b] (K) .29[b] (1st) .37[c] (K) .07[c] (1st .13[g] (1st)			
Kaminski & Good (1996)	Grades K–1 (N = 78) 99% C, 1% O 12% FCL						.83–.93[h] .46[i] .02–.65[j] .55[k] .29–.55[l] .43[m]				
Riedel (2007)	Grade 1 (N = 1518) 92% AA, 8% O 85% FCL; 4% ELL						.15–2.6[1n] .19–.31[2n] 18–.26[o]		56[n] 54[o]	56[n] 54[o]	
Schilling et al. (2007)	Grade 1 (N = 2,588) 60% AA, 25% C, 13% H, 2% O 81% FCL						.32[p]	.40–.43[p]			

(Table 3 continues)

Table 3 continued

Authors	Sample	Reliability Evidence			Validity Evidence			Decision Making (%)			
		Test–Retest	Alt. Form	Inter-Rater	Convergent	Concurrent	Predictive	Sens.	Spec.	Pos.	Neg.
				Dissertations and Theses							
Cook (2003)	Grade 1 (N = 79) 100% Caucasian						.18–.54[q]				
Fien (2004)	Grade K (N = 3652) Grade 1 (N = 3501)		.62								
Fleming (1999)	Grade K (N = 82)						.40[t] .36[g]				
Greene (2002)	Grade Pre-K (N = 25)							−.04–.52[t]			

Note. PSF = Phonemic-Segmentation Fluency; Alt. = alternate; Sens. = sensitivity; Spec. = specificity; C = Caucasian; H = Hispanic; O = Other; FCL = Eligible for a free or reduced-cost lunch; ELL = English language learners; CBM = Curriculum-Based Measurement; CTOPP = Comprehensive Test of Phonological Processing; DIBELS = Dynamic Indicators of Basic Early Literacy Skills.

[a] Test of Word Reading Efficiency.
[b] DIBELS Nonsense Word Fluency.
[c] CBM Oral Reading Fluency.
Reliability, Validity, and Decision-Making Evidence for PSF
[d] CTOPP Phonological Awareness Composite.
[e] CTOPP Phonological Memory Composite.
[f] CTOPP Rapid Naming Composite.
[g] DIBELS Oral Reading Fluency.
[h] McCarthy Scales of Children's Abilities.
[i] Metropolitan Readiness Test.
[j] Rhode Island Pupil Identification Scale.
[k] Curriculum-Based Measurement.
[l] Teacher Rating Scale.
[m] Stanford Diagnostic Reading Test.
[1n] Group Reading Assessment and Diagnostic Evaluation (English-Speaking Students).
[2n] Group Reading Assessment and Diagnostic Evaluation (ELL Students).
[o] TerraNova CAT Reading.
[p] Iowa Test of Basic Skills (Reading Composite).
[q] Stanford Achievement Test, 9th Edition (Reading).
[r] Rapid Automatized Naming.
[s] Wechsler Preschool and Primary Scale of Intelligence—Revised (Verbal).
[t] Early Screening Profile.
[u] Slosson Oral Reading Test.

Table 4
Reliability, Validity, and Decision-Making Evidence for NWF

		Reliability Evidence				Validity Evidence			Decision Making (%)			
Authors	Sample	Test–Retest	Alt. Form	Intern Cons.	Inter-Rater	Convergent	Concurrent	Predictive	Sens.	Spec.	Pos.	Neg.
					Peer Reviewed							
Burke & Hagan-Burke (2007)	Grade 1 (N = 213) 53% C, 24% AA, 6% H, 17% O 33% FCL						.68–.75[a]					
Fuchs & Fuchs (1999)	Grade 1 (N = 151) 38% AA, 35% C, 24% H 50% FCL; 7% SPLED						.50–.64[b] .54–.80[c]	.46–.57[b] .64–.80[c] .66[j]				
Good et al. (2001)	Grades K–3 (N = 302–378) 90% C, 10% O 37–63% FCL							.78[d]	.91[d]	90[d]		
Jordan et al. (2006)	Grade K (N = 411) 43% C, 36% AA, 15% H, 6% O	.92										
Kamii & Manning (2005)	Grades K–1 (N = 208) 85–95% C, 4–10% AA, 1–5% O							.56[k] (K) .62[k] (1st) .70[d] (1st)				
Kamps et al. (2003)	Grades K–2 (N = 383) 40% AA, 34% C, 8% H, 18% O							.79[e]	.78[d]			
McMaster et al. (2005)	Grade 1 (N = 176)	.87						.51–.65[b] .50–.80[c]				
Riedel (2007)	Grade 1 (N = 1518) 92% AA, 8% O 85% FCL; 4% ELL							.45–.46[f] non .41–.47[f] ELL .37–.39[g]		68[f] 62[g]	65[f] 58[g]	

(Table 4 continues)

Table 4 continued

Authors	Sample	Reliability Evidence				Validity Evidence			Decision Making (%)			
		Test–Retest	Alt. Form	Intern Cons.	Inter-Rater	Convergent	Concurrent	Predictive	Sens.	Spec.	Pos.	Neg.
Schilling et al. (2007)	Grade 1 (N = 2,588) 60% AA, 25% C, 13% H, 2% O 81% FCL						.60[h]	.57–.60[h]				
				Dissertations and Theses								
Cook (2003)	Grade 1 (N = 79); 100% C							.57–.64[i]				
Fien (2004)	Grade K (N = 3652); Grade 1 (N = 3501)		.58									

Note. Studies were peer reviewed. NWF = Nonsense Word Fluency; Alt. = alternate; Sens. = Sensitivity; Spec. = specificity; Pos. = positive predictive power; Neg. = negative predictive power; C = Caucasian; AA = African American; H = Hispanic; O = other; FCL = free or reduced-cost lunch; SPLED = Special Education; ELL = English language learners; DIBELS = Dynamic Indicators of Basic Early Literacy.

[a] Test of Word Reading Efficiency.
[b] Woodcock Reading Mastery Test—Revised.
[c] Comprehensive Reading Assessment Battery (Fluency).
[d] DIBELS Oral Reading Fluency.
[e] DIBELS Letter Naming Fluency.
[f] Group Reading Assessment and Diagnostic Evaluation.
[g] TerraNova CAT Reading.
[h] Iowa Test of Basic Skills (Reading Composite).
[i] Stanford Achievement Test, 9th Edition (Reading).
[j] Comprehensive Reading Assessment Battery (Comprehension).
[k] Slosson Oral Reading Test.

efficients, all of which were adequate for screening purposes. However, no reviewed studies addressed interrater reliability evidence for NWF. Evidence of convergent validity also was not addressed in any reviewed studies. Seven peer-reviewed articles reported concurrent validity evidence across 11 different literacy measures, with coefficients ranging from moderate to high (*Mdn* = .58). Predictive validity evidence across three studies also ranged from moderate to high (*Mdn* = .62).

Accuracy and diversity considerations. Two peer-reviewed studies reported decision-making accuracy statistics for NWF. Evidence of sensitivity and specificity was adequate for both studies. Positive predictive power and negative predictive power, however, were not addressed in any of these studies. Similar to LNF and PSF, Riedel (2007) reported data for students of minority status, and coefficients did not differ from studies consisting of primarily Caucasian samples.

ORF

Reliability and validity evidence. Finally, psychometric evidence for ORF is presented in Table 5. Evidence of reliability for ORF was reported in three peer-reviewed articles and one dissertation. Test–retest reliability, reported in one study, was adequate for screening and group decision-making purposes. Two studies also addressed alternate-form reliability, whereas only one article reported inter-rater reliability coefficients. Both alternate-form and interrater reliability evidence exceeded group decision-making standards, and all of the reported coefficients exceeded .80.

With regard to validity, one technical report included convergent validity evidence for ORF. Concurrent validity evidence ranged from moderate to high (*Mdn* = .71) across seven peer-reviewed articles, two dissertations, and five technical reports. Finally, three peer-reviewed articles reported predictive validity evidence. These values were within the moderate to high range (*Mdn* = .68) across studies. ORF scores demonstrated the strongest reliability and criterion-related validity evidence across multiple peer-reviewed studies, dissertation/theses, and technical reports.

Accuracy and diversity considerations. Three peer-reviewed articles and five technical reports included decision-making accuracy statistics for ORF with statewide standardized achievement measures. Sensitivity (*Mdn* = 77%) and specificity (*Mdn* = 88%) values reported in these studies were adequate for screening purposes.

Diverse racial and ethnic samples also were more prevalent for ORF studies, with participants of Hispanic, African American, and Latino descent represented. As with previous indicators, only one study (Riedel, 2007) reported separate coefficients for students of minority status. This study included African American students, and coefficients did not differ from studies consisting of primarily Caucasian samples.

Discussion

The purpose of this study was to systematically review and synthesize published reliability and validity evidence for DIBELS. Searches of multiple electronic databases ultimately yielded 27 studies that met criteria for inclusion in the review. Results indicated that the amount and quality of evidence differed by indicator, with several gaps in the empirical literature. These results are discussed relative to each of the guiding research questions.

Research Question 1: Reliability of DIBELS Scores for Screening and Group Decisions

Several indicators have strong evidence of score reliability for screening and group decision-making purposes. Evidence for LNF and ORF is particularly robust across multiple indices (i.e., test–retest, alternate form, and inter-rater), suggesting that both probes can be considered consistent measures of student performance across time periods, forms, and examiners. Although published reliability evidence is satisfactory overall for PSF and NWF, no inter-rater reliability studies have

been published to date. Given multiple educational professionals (psychologists, teachers, specialists) may play a role in DIBELS's administration over the course of a school year, further evidence is needed to determine the reliability of students' scores across examiners of various educational backgrounds and assessment experience levels. Finally, published evidence for ISF was consistently sparse across test–retest, alternate-form, and interrater reliability. Thus, no firm conclusions can be drawn regarding the adequacy of DIBELS ISF scores, even for screening purposes.

Research Question 2: Criterion-Related and Predictive Validity Evidence

Evidence of score validity for single-point decision making is relatively promising across DIBELS indicators, although several gaps exist. Not surprisingly, ORF scores demonstrated excellent concurrent and predictive validity evidence across multiple studies. Two indicators (i.e., LNF, NWF) lack convergent validity evidence, and such coefficients were only reported in one study for ISF and PSF. Although concurrent validity evidence is more widely available for NWF and PSF, the variable magnitude of coefficients (relatively small in some studies) indicates that these measures should be interpreted with caution. Similarly, predictive validity coefficients for NWF and PSF were fairly low and only reported in a limited number of studies. Both concurrent and predictive validity evidence is more promising for LNF, ranging from the moderate to high across studies. Finally, no evidence of predictive validity was available for ISF; concurrent and convergent validity support was relatively weak. In sum, ORF and, to a slightly lesser extent LNF, have the strongest validity evidence base, but the remaining indicators would benefit from additional validity studies.

Research Question 3: Diagnostic Accuracy for Screening Purposes

In the current review, decision-making accuracy evidence varied across indicators. For ISF and PSF, sensitivity was adequate and specificity was relatively low across multiple studies. Conversely, positive predictive power was low and negative predictive power was high. Decision-making evidence for LNF demonstrated a similar pattern in one study (Hintze et al., 2003), but both sensitivity and specificity levels were low in another (Riedel, 2007). Evidence for NWF also was mixed, with one study indicating high levels of sensitivity and specificity for this indicator and a second study indicating lower levels. No studies included predictive power indices for NWF. With the exception of two studies, ORF exhibited high levels of sensitivity and specificity. One study (Roehrig, Petscher, Nettles, Hudson, & Torgesen, 2008) reported predictive power for ORF, and similar to the other DIBELS, results indicated high negative predictive power and lower positive predictive power.

Within a response to intervention model of service delivery, accurate intervention decisions are necessary to ensure optimal use of available resources. Results of the current review suggest that ORF scores are the most accurate in predicting future literacy proficiency, although all DIBELS indicators may overidentify students as at risk. Thus, practitioners should consider additional data to supplement screening decisions based on DIBELS, particularly those decisions based on the four indicators other than ORF.

Research Question 4: Variability Across Racial and Ethnic Groups

Two of the reviewed studies (Clarke et al., 2003; Riedel, 2007) featured samples with a majority (≥80%) of participants from racial and ethnic minority groups. Several other studies (e.g., Burke & Hagan-Burke, 2007; Kamps et al., 2003; Francis et al., 2008; Schilling, Carlisle, Scott, & Zeng 2007) included samples that were quite heterogeneous, which oversampled students of minority status relative to the demographic characteristics of the U.S. school-age population. Although psychometric evidence did not appear to differ substantially between studies featuring racially and ethnically diverse samples from

Table 5
Reliability, Validity, and Decision-Making Evidence for ORF

Authors	Sample	Reliability Evidence			Validity Evidence			Decision Making (%)			
		Test–Retest	Alt. Form	Inter-Rater	Convergent	Concurrent	Predictive	Sens.	Spec.	Pos.	Neg.
					Peer Reviewed						
Burke & Hagan-Burke (2007)	Grade 1 (N = 213) 53% C, 24% AA, 6% H, 17% O 33% FCL						.77–.92[a]				
Francis et al. (2008)	Grade 2 (N = 134) 57% H, 31% AA, 9% C, 3% O 80% FCL		.87–.93	.85		.69–.88[b]					
Graves, Gersten, & Haager (2004)	Grades 1 (N = 186) 100% FCL; 80–100% ELL					.65[c]					
Good et al. (2001)	Grades K–3 (N = 302–378) 90% C, 10% O 37–63% FCL	.82					.67[g]	72%[h]	96%[h]		
Kamps et al. (2003)	Grades K–2 (N = 383) 40% AA, 34% C, 8% H, 18% O					.78[d] .74[e]					
Riedel (2007)	Grade 1 (N = 1518) 92% AA, 8% O 85% FCL; 4% ELL					.59–.67[1h] .72–.80[2h] .49–.54[i]	.77[h] .69[i]	76%[h] 65[i]			
Roehrig et al. (2008)	Grade 3 (N = 35,207) 36% C, 36% AA, 23% H, 5% O 75% FCL					.70–.71[j] .70–.71[k]	.66–.68[j] .68–.69[k]	88[j]	80[j]	56[j]	96[j]
Schilling et al. (2007)	Grade 1 (N = 2,588) 60% AA, 25% C, 13% H, 2% O 81% FCL					.75[l]	.69[l]				
					Dissertations and Theses						
Cook (2003)	Grade 1 (N = 79) 100% C					.61–.75[m]					
Fien (2004)	Grade K (N = 3652) Grade 1 (N = 3501)		.85								
Greene (2002)	Grade Pre-K (N = 25)					.37[n]					

Table 5 continued

Authors	Sample	Reliability Evidence			Validity Evidence			Decision Making (%)			
		Test–Retest	Alt. Form	Inter-Rater	Convergent	Concurrent	Predictive	Sens.	Spec.	Pos.	Neg.
					Technical Reports						
Barger (2003)	Grade 3 (N = 38)					.73[o]		66[o]	100[o]		
Buck & Torgesen (2003)	Grade 3 (N = 1102) 83% C, 7% AA, 6% H, 4% O 46% FCL; 1% ELL					.70[j]			77[j]	92[j]	
Good, Kaminski, Smith & Bratten (2001)	Not reported		.91–.96		.91–.96[f]						
Shaw & Shaw (2002)	Grade 3 (N = 58)					.73–.80[p]			43[p]	90[p]	
Vander Meer, Lentz, & Stollar (2005)	Grades 3–4 (N = 364) 14% FCL					.61–.65[q]			96[q]	72[q]	
Wilson (2005)	Grade 3 (N = 241) .59% H, 34% C, 7% O 71% FCL					.74[r]			93[r]	82[r]	

Note. ORF = Oral Reading Fluency; Alt. = alternate; Sens. = sensitivity; Spec. = specificity; Pos. = positive predictive power; Neg. = negative predictive power; C = Caucasian; AA = African American; H = Hispanic; O = Other; FCL = free or reduced-cost lunch; ELL = English language learners; DIBELS = Dynamic Indicators of Basic Early Literacy.
[a] Test of Word Reading Efficiency.
[b] The Primary Reading Inventory.
[c] English Language Learners Classroom Observation Instrument.
[d] DIBELS Nonsense Word Fluency.
[e] DIBELS Letter Naming Fluency.
[f] Tests of Reading Fluency.
[g] Oregon Statewide Assessment.
[1h] Group Reading Assessment and Diagnostic Evaluation (English-Speaking Students).
[2h] Group Reading Assessment and Diagnostic Evaluation (ELL Students).
[i] TerraNova CAT Reading.
[j] Florida Comprehensive Assessment Test: Sunshine State Standards (Reading).
[k] Stanford Achievement Test, Tenth Edition (SAT-10).
[l] Iowa Test of Basic Skills (Reading Composite).
[m] Stanford Achievement Test, 9th Edition (Reading).
[n] Early Screening Profile.
[o] North Carolina End of Grade Reading Assessment.
[p] Colorado State Assessment Program.
[q] Ohio Fourth Grade Reading Proficiency Test.
[r] Arizona Instrument to Measure Standards.

those that used primarily Caucasian samples, it is very difficult to draw any firm conclusions from the current review for two reasons. First, few criterion measures were consistent across studies. Second, only Riedel (2007) directly compared evidence across groups, and these comparisons were solely based on language (i.e., English language learners vs. native English speakers). Thus, additional studies are necessary to address the technical adequacy of DIBELS scores for students from specific racial and ethnic groups, as well as English language learners.

Directions for Future Research

As noted previously, there are some common evidence limitations across all of the indicators that need to be addressed in future studies. Specifically, future research is necessary to strengthen the reliability and validity evidence base for at least four of the five DIBELS indicators. Such studies should examine the forms of evidence outlined by AERA, APA, and NCME (1999), with particular consideration to the examination of variation in evidence across race, socioeconomic status, and levels of English language proficiency.

There is a paucity of DIBELS studies available for Stages 2 and 3 of Fuchs' (2004) three-stage framework of research evidence for progress monitoring measures. More specifically, the studies examined the technical features (i.e., reliability and validity) of single-point scores. Thus, further studies are needed to determine technical features of slopes resulting from the repeated administration of DIBELS for progress-monitoring purposes. One important question is determining whether changes in general outcome measure scores over time are reflective of changes in the overarching academic skill (general outcome) of interest (Fuchs). Although no studies identified during our search directly addressed such questions, related studies have begun to appear recently in the literature (e.g., Ardoin & Christ, 2008; Baker et al., 2008; Good, Baker, & Peyton, 2009).

There are a variety of reasons why change on progress monitoring measures, such as DIBELS, may not correspond with change in the general academic skill of interest. Practice effects, for example, could lead to improved performance on a standardized measurement task even though students' overall skills are not improving. Hintze and Christ (2004) and Christ (2006) have emphasized the importance of considering the standard error of the slope (*SEb*) in the context of progress monitoring decisions. Few researchers have explored slope stability, or more specifically, factors that affect slope stability (e.g., testing conditions, inconsistent probe skill levels, practice effects), for families of general outcome measures such as DIBELS.

Future studies also are necessary to examine the effects of DIBELS use on classroom instruction, and ultimately, student achievement. Several such studies have been completed with CBM for students with and without identified disabilities (Stecker, Fuchs, & Fuchs, 2005), but further studies evaluating instructional outcomes resulting from the use of DIBELS probes are needed.

Recommendations for Practice

Given the available reliability and validity evidence, caution should be exercised when incorporating ISF into either screening or progress-monitoring efforts in the schools. To date, the effectiveness of this measure to differentiate student skills, as well as to predict future outcomes on standardized statewide achievement measures, is unknown. Until future evidence suggests otherwise, ISF would be most appropriately used in conjunction with other assessment data to inform instructional practices by gauging students' development of skills in identifying initial sounds.

LNF appears to be a promising measure for both screening and progress monitoring purposes because the scores appear to be stable across time periods, forms, and examiners. A variety of published studies have demonstrated strong relationships between LNF scores and future scores on a variety of reading outcome measures. However, decision-making accuracy is varied and limited for this

measure, which suggests that screening scores should not be used in isolation for individual decision-making purposes. Instead, failure to meet the predetermined grade-level benchmark should necessitate administration of a follow-up assessment battery with high established levels of decision-making accuracy (i.e., sensitivity, specificity, positive predictive power, and negative predictive power). Moreover, as with all other indicators, further evidence on target slope levels is needed.

Stability of scores for PSF and NWF is promising for progress monitoring and screening purposes. However, lack of convergent validity and low levels of concurrent and predictive validity potentially limit the utility of scores from these measures. None of the reviewed studies reported positive predictive power or negative predictive power values for NWF, whereas only one study was available for each of the remaining indicators. Given the importance of positive predictive power and negative predictive power to evaluating classification accuracy (Hambleton et al., 1978; Livingston & Lewis, 1995) and the use of DIBELS benchmark assessments for identifying children in need of intervention, the apparent paucity of such data is a significant limitation. Thus, although these indicators may provide more reliable information than ISF, they should ideally be considered in conjunction with additional assessment data.

Evidence supports the validity of ORF as a screening and progress monitoring tool across the primary grade levels (i.e., beginning at the Grade 1 winter benchmark), which suggests that practitioners should weigh the additive contributions of DIBELS indicators other than ORF for both screening and progress monitoring decisions. Furthermore, utilizing measures that may inaccurately identify students as at risk could result in misallocation of limited district time and resources from an "educational cost-benefit analysis perspective" (Hintze et al., 2003, p. 555).

Conclusion

Although the extant reliability and validity evidence of DIBELS scores for single-point decisions is promising, there are several psychometric gaps that remain to be addressed. Most important, future studies should validate grade-based target slope levels for DIBELS performance (e.g., Christ, 2006; Fuchs, 2004; Hintze & Christ, 2004). The effectiveness of current DIBELS cut scores should be investigated across indicators, grade levels, and diverse racial and ethnic groups. Finally, the limited ability of Grade 1 early literacy indicators (i.e., LNF, PSF, NWF) to predict future academic achievement beyond ORF performance should be carefully considered by practitioners. DIBELS has emerged as one of the most frequently utilized assessment measures of the early 21st century. Although published empirical support for most indicators is adequate, evidence to justify use of these measures for progress monitoring and screening decisions should be further examined to ensure that we are making the best possible decisions for our emerging and early readers.

Footnotes

*References marked with an asterisk indicate studies that were included in the review.

References

Adams, M. J. (1990). *Beginning to read: Thinking and learning about print.* Cambridge, MA: MIT Press.

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

Ardoin, S. P., & Christ, T. J. (2008). Evaluating curriculum-based measurement slope estimates using data from triannual universal screenings. *School Psychology Review, 37,* 109–125.

Baker, S. K., Smolkowski, K., Katz, R., Fien, H., Seeley, J., Kame'enui, E., et al. (2008). Reading fluency as a predictor of reading proficiency in low-performing high poverty schools. *School Psychology Review, 37,* 18–37.

*Barger, J. (2003). *Comparing the DIBELS Oral Reading Fluency indicator and the North Carolina End of Grade Reading Assessment* (Technical Report). Asheville: North Carolina Teacher Academy.

Blachman, B. A. (1991). Phonological awareness: Implications for prereading and early reading instruction. In S. A. Brady & D. P. Shankweiler (Eds.), *Phonological processes in literacy* (pp. 29–36). Hillsdale, NJ: Erlbaum Associates.

Brunsman, B. A. (2003). Review of the DIBELS: Dynamic Indicators of Basic Early Literacy Skills, Sixth

Edition. In B. S. Plake, J. C. Impara, & R. A. Spies (Eds.), *The fifteenth mental measurements Yearbook* (pp. 307–310). Lincoln, NE: Buros Institute of Mental Measurements.

*Buck, J., & Torgesen, J. (2003). *The relationship between performance on a measure of oral reading fluency and performance on the Florida Comprehensive Assessment Test* (FCRR Technical Report No. 1). Tallahassee: Florida Center for Reading Research.

*Burke, M. D., & Hagan-Burke, S. (2007). Concurrent criterion-related validity of early literacy indicators for middle of first grade. *Assessment for Effective Intervention, 32,* 66–77.

Byrne, B., & Fielding-Barnsley, R. (1989). Phonemic awareness and letter knowledge in the child's acquisition of alphabetic principle. *Journal of Educational Review, 81,* 313–321.

Christ, T. J. (2006). Short-term estimates of growth using curriculum-based measurement of oral reading fluency: Estimating standard error of the slope to construct confidence intervals. *School Psychology Review, 35,* 128–133.

*Clarke, A. T., Power, T. J., Blom-Hoffman, J., Dwyer, J. F., Kelleher, C. R., & Novak, M. (2003). Kindergarten reading engagement: An investigation of teacher ratings. *Journal of Applied School Psychology, 20,* 131–144.

*Cook, R. G. (2003). *The utility of DIBELS as a curriculum based measurement in relation to reading proficiency on high stakes tests.* Unpublished master's thesis, Marshall University Graduate College.

Cunningham, A. E., & Stanovich, K. E. (1998). The impact of print exposure on word recognition. In J. L. Metsala, & L. C. Ehri (Eds.), *Word recognition in beginning literacy* (pp. 235–262). Mahwah, NJ: Lawrence Erlbaum Associates Publishers.

*Elliott, J., Lee, S. W., & Tollefson, N. (2001). A reliability and validity study of the Dynamic Indicators of Basic Early Literacy Skills—Modified. *School Psychology Review, 30(1),* 33–49.

*Fien, H. (2004). *An examination of school and individual student level predictors of successful reading and reading related outcomes for kindergarten and first grade outcomes: A comparison of two models of school-wide reading reform.* Unpublished doctoral dissertation, University of Oregon.

*Fleming, K. M. (1999). *The effect of instruction, rapid automatized naming, intellectual functioning, initial phonological awareness skill, and age on phonological awareness growth trajectories.* Unpublished doctoral dissertation, University of Oregon, Eugene.

*Francis, D. J., Santi, K. L., Barr, C., Fletcher, J. M., Varisco, A., & Foorman, B. R. (2008). Form effects on the estimation of students' oral reading fluency using DIBELS. *Journal of School Psychology, 46,* 315–342.

Fuchs, L. S. (2004). The past, present, and future of curriculum-based measurement research. *School Psychology Review, 33,* 188–192.

*Fuchs, L. S., & Fuchs, D. (1999). Monitoring early reading development in first grade: Word Identification Fluency versus Nonsense Word Fluency. *Exceptional Children, 71,* 7–21.

Good, R. H., Baker, S. K., & Peyton, J. A. (2009). Making sense of Nonsense Word Fluency: Determing adequate progress in early first-grade reading. *Reading & Writing Quarterly, 25,* 33–56.

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 at 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). Eugene: University of Oregon.

Good, R. H., Kaminski, R. A., Smith, S., Simmons, D., Kame'enui, E., & Wallin, J. (2003). Reviewing outcomes: Using DIBELS to evaluate kindergarten curricula and interventions. S. R. Vaughn & K. L. Briggs (Eds.), *Reading in the classroom: Systems for the observation of teaching and learning* (pp. 221–259). Baltimore: Brooks.

*Good, R. H., Simmons, D., & 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.

*Graves, A. W., Gersten, R., & Haager, D. (2004). Literacy instruction in multiple-language first-grade classrooms: Linking students outcomes to observed instructional practice. *Learning Disabilities Research & Practice, 19,* 262–272.

*Greene, L. S. (2002). *Investigating parent-child storybook reading and its relationship to early literacy skills: Development and use of direct observation system.* Unpublished doctoral dissertation, University of Massachusetts Amherst.

Hambleton, R. K., Swaminathan, H., Algina, J., & Coulson, D. B. (1978). Criterion-referenced testing and measurement: A review of technical issues and developments. *Review of Educational Research, 48,* 1–47.

Hasbrouck (1998). *Reading fluency: Principles for instruction and progress monitoring* (Professional Development Guide). Austin: Texas Center for Reading and Language Arts, University of Texas at Austin.

Hintze, J. M., & Christ, T. J. (2004). An examination of variability as a function of passage variance in CBM progress monitoring. *School Psychology Review, 33,* 204–217.

Hintze, J. M., Christ, T. J., & Methe, S. A. (2006). Curriculum-based assessment. *Psychology in the Schools, 43,* 45–56.

*Hintze, J. M., Ryan, A. L., & Stoner, G. (2003). Concurrent validity and diagnostic accuracy of the Dynamic Indicators of Basic Early Literacy Skills (DIBELS) and the Comprehensive Test of Phonological Processing. *School Psychology Review, 32(4),* 541–556.

*Jordan, N. C., Kaplan, D., Oláh, L. N., & Locuniak, M. N. (2006). Number sense growth in kindergarten: A longitudinal investigation of children at risk for mathematics difficulties. *Child Development, 77,* 153–175.

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

*Kamii, C., & Manning, M. (2005). Dynamic Indicators of Basic Early Literacy Skills (DIBELS): A tool for evaluating student learning? *Journal of Research in Childhood Education, 20,* 75–90.

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

*Kamps, D. M., Willis, H. P., Greenwood, C. R., Thorne, S., Lazo, J. F., Crocket, J. L., et al. (2003). Curriculum influences on growth in early reading fluency for students with academic and behavioral risks: A descriptive study. *Journal of Emotional and Behavioral Disorders, 11,* 211–224.

Livingston, S. A., & Lewis, C. (1995). Estimating the consistency and accuracy of classifications based on test scores. *Journal of Educational Measurement, 32,* 179–197.

Lundberg, I., Frost, J., & Petersen, O. P. (1988). Effects of an extensive program for stimulating phonological awareness in preschool children. *Reading Research Quarterly, 23,* 263–284.

*McMaster, K. L., Fuchs, D., Fuchs, L. S., & Compton, D. L. (2005). Responding to nonresponders: An experimental field trial of identification and intervention methods. *Exceptional Children, 71,* 445–464.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50,* 741–749.

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 Academics Press.

No Child Left Behind Act of 2001, Pub. L. No. 107–110, 2204 (2002).

*Riedel, B. W. (2007). The relation between DIBELS, reading comprehension, and vocabulary in urban first-grade students. *Reading Research Quarterly, 42,* 546–567.

*Roehrig, A. D., Petscher, Y., Nettles, S. M., Hudson, R. F., & Torgesen, J. K. (2008). Accuracy of the DIBELS oral reading fluency measure for predicting third grade reading comprehension outcomes. *Journal of School Psychology, 46,* 343–366.

Salvia, J., Ysseldyke, J., & Bolt, S. (2010). *Assessment in special and inclusive education* (12th ed.). Florence, KY: Wadsworth Publishing.

Samuels, S. J. (2007). The DIBELS tests: Is speed of barking at print what we mean by reading fluency? *Reading Research Quarterly, 42,* 563–566.

*Schilling, S. G., Carlisle, J. F., Scott, S. E., & Zeng, J. (2007). Are fluency measures accurate predictors of reading achievement? *The Elementary School Journal, 107,* 429–448.

*Shaw, R., & Shaw, D. (2002). *DIBELS Oral Reading Fluency-based indicators of third grade reading skills for Colorado State Assessment Program (CSAP)* (Technical Report). Eugene: University of Oregon.

Stanovich, K. E. (1986). Matthew effects in reading: Some consequences of individual differences in the acquisition of literacy. *Reading Research Quarterly, 21,* 360–406.

Stecker, P. M., Fuchs, L. S., & Fuchs, D. (2005). Using curriculum-based measurement to improve student achievement: Review of research [Special Issue]. *Psychology in the Schools, 42,* 795–819.

Swets, J. A. (1988). Measuring the accuracy of diagnostic systems. *Science, 24*, 1285–1293.

Torgesen, J. K. (2002). The prevention of reading difficulties. *Journal of School Psychology, 40,* 7–26.

University of Oregon Center on Teaching and Learning. (2009). *Dynamic Indicators of Basic Early Literacy Skills*. Retrieved January 10, 2009, from http://dibels.uoregon.edu/

*Vander Meer, C. D., Lentz, F. E., & Stollar, S. (2005). *The relationship between oral reading fluency and Ohio proficiency testing in reading* (Technical Report). Eugene: University of Oregon.

Wayman, M. M., Wallace, T., Wiley, H. I., Ticha, R., & Espin, C. A. (2007). Literature synthesis on curriculum-based measurement in reading. *The Journal of Special Education, 41,* 85–120.

Wilson, J. (2005). The relationship of Dynamic Indicators of Basic Early Literacy Skills Oral Reading Fluency to performance on Arizona Instrument to Measure Standards* (Technical Report). Tempe, AZ: Tempe School District.

Date Received: January 12, 2009
Date Accepted: March 14, 2010
Action Editor: Matthew Burns ■

Catherine T. Goffreda, PhD, is a recent graduate from the school psychology program at The Pennsylvania State University. She completed her predoctoral internship in the School District of Broward County, Florida and is a school psychologist for Guilford County schools in Greensboro, NC.

James Clyde DiPerna, PhD, is Associate Professor and Harry and Marion Eberly Faculty Fellow in the School Psychology Program at The Pennsylvania State University. His research focuses on assessment and intervention strategies to promote students' academic, social, and emotional competence.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.