School Psychology Quarterly
2008, Vol. 23, No. 4, 542–552

1045-3830/08/$12.00 DOI: 10.1037/a0013245

Beyond Correlational Analysis of the Dynamic Indicators of Basic Early Literacy Skills (DIBELS): A Classification Validity Study

Jason M. Nelson
University of Georgia

This study investigated the classification validity of the Dynamic Indicators of Basic Early Literacy Skills (DIBELS) using a sample of kindergarteners ($N = 177$). Results indicated the cutoff scores for determining *at-risk* status on the DIBELS produced substantial false negative rates. Cutoff scores identifying students as at *some risk* produced substantial false positive rates. At both levels of risk status, the DIBELS showed low positive predictive power, but high negative predictive power, indicating it was far better at identifying students with adequate reading skills than those with inadequate reading skills. Recommendations for appropriate use of the DIBELS for reading screening and suggestions for future research are provided.

Keywords: DIBELS, early literacy, reading screening, diagnostic accuracy, prevention of reading disabilities

Of all academic problems that lead to placement in special education, difficulty with learning how to read is most pervasive. Nearly 80% of referrals for special education evaluations involve reading problems (Nelson & Machek, 2007). Fortunately, recent advances in reading research have indicated that this current state of affairs is alterable. The clearest and least controversial solution is for schools to attempt to prevent reading difficulties from developing (Torgesen, 2002). As a first step in this preventive approach, at-risk readers must be identified as early and accurately as possible. To meet the needs of prevention-oriented educational service delivery models (e.g., response to intervention [RtI] models), several early reading screening instruments have been created. One early reading screening instrument, the Dynamic Indicators of Basic Early Literacy Skills (DIBELS; Good & Kaminski, 2002), has, according to Manzo (2005), "come to symbolize the standard for early literacy assessment throughout much of the country" (p. 1).

Upon completing the most exhaustive review of early reading instruments to date, the Reading First Assessment Committee (RFAC; Kame'enui et al., 2002) concluded that the DIBELS was a tool with sufficient scientific evidence for use as a reading screener. Despite this conclusion, the RFAC described the overall findings of their review of the technical adequacy of early reading measures as both "disturbing" (Kame'enui et al., 2006, p. 7) and "sobering" (p. 9). No reading instruments met all evaluation criteria. The RFAC's review indicated that validity evidence for nearly all of the instruments was derived from correlational data (i.e., criterion validity), and the classification validity of the majority of instruments (including the DIBELS) had not been investigated. The study of classification validity is regarded as "the sine qua non of screening research" (Jenkins, 2003, p. 6) and more adequate than investigation of criterion validity in determining the utility of reading screening measures (Bishop, 2003).

That an instrument is correlated with various criterion measures only provides weak evidence for its utility as a screener (Jenkins, 2003). Rather than calculating correlations, classification validity is examined by comparing the number of individuals identified as exhibiting (and not exhibiting) problems on a "gold standard" test (i.e., true positives and negatives) as compared with those identified as at risk on a screening test. The gold standard (also referred to as the reference standard) is considered to be the best available evidence for the existence of

Jason M. Nelson, Regents Center for Learning Disorders, University of Georgia.

Correspondence concerning this article should be addressed to Jason M. Nelson, Regents' Center for Learning Disorders, University of Georgia, 337 Milledge Hall, Athens, GA 30602. E-mail: jmnelson@uga.edu

a particular condition or characteristic (Kessel & Zimmerman, 1993).

Since the publication of the RFAC report in 2002, one study (Hintze, Ryan, & Stoner, 2003) on the classification validity of the DIBELS has been published in a peer-reviewed journal. Hintze et al. found that the cutoff scores for determining risk status on the DIBELS recommended by the test authors produced generally poor diagnostic accuracy characteristics. Consequently, Hintze et al. examined a range of cutoff scores on the DIBELS to determine a level of diagnostic accuracy they deemed acceptable. These analyses resulted in different cutoff scores than those recommended by the developers of the DIBELS.

Purpose of Study

The purpose of the current study was to extend investigation of the classification validity of the DIBELS. Diagnostic accuracy of the DIBELS was investigated in two ways. First, the DIBELS and a norm-referenced test of phonological awareness were administered concurrently to a group of kindergarteners at midyear. Second, the same students were administered a norm-referenced test of reading skill at the end of their kindergarten year.

Research Questions

Four research goals were pursued. The first was to determine the diagnostic accuracy characteristics of the individual DIBELS tasks when using the cutoff scores for determining risk status recommended by the test developers. Second, alternative cutoff scores to those recommended by the test developers were examined to determine those needed to attain specific diagnostic accuracy characteristics. Like Hintze et al. (2003), a range of cutoff scores was explored to determine those that produced both sensitivity and specificity of 75%. Additionally, cutoff scores were explored to determine those that produced the highest level of specificity while maintaining at least 90% sensitivity. The rationale for this analysis was that many DIBELS users are likely interested in identifying as many at-risk readers as possible, while concurrently accurately ruling out those who are not at risk. Third, the *overall* diagnostic accuracy of each DIBELS tasks was examined, and the tasks were compared to determine if some were superior to others. Finally, the diagnostic accuracy characteristics of the DIBELS tasks *together* rather than *independently* were explored.

Method

Participants

Participants were 177 kindergarten students enrolled in a small city public school system in a Midwestern state. They were recruited from 10 classrooms and 2 schools. Participants' average age was 5.44 years ($SD = .50$). The sample's ethnicity breakdown was as follows: White (92.7%; $n = 164$), African American (2.3%; $n = 4$), Hispanic (4.0%; $n = 7$), Asian American (.6%; $n = 1$), and other (.6%; $n = 1$).

Instruments

DIBELS

The DIBELS Initial Sound Fluency (ISF), Letter Naming Fluency (LNF), Phoneme Segmentation Fluency (PSF), and Nonsense Word Fluency (NWF) tasks were administered. Published studies on the DIBELS (e.g., Kaminski & Good, 1996) have indicated acceptable levels of reliability measures, including test-retest, alternate forms, and inter-rater. Several studies (e.g., Hintze et al., 2003) have provided support for the criterion-related validity of the DIBELS.

Three levels of risk status—at risk, some risk, and low risk—are designated by the test developers (see *DIBELS Benchmark Goals,* n.d.). Students are potentially at risk for reading difficulty at midkindergarten if they score below 10 on ISF, 15 on LNF, 7 on PSF, or 5 on NWF. Performance from 10 to 24 on ISF, 15 to 26 on LNF, 7 to 17 on PSF, and 5 to 12 on NWF places the student at some risk for later reading problems. Scores above the high end of these ranges are suggestive of low risk status.

Test of Phonological Awareness – Second Edition: Plus

The Test of Phonological Awareness – Second Edition: Plus (TOPA-2+; Torgesen & Bryant, 2004a) is a norm-referenced test that measures phonological awareness and knowledge of letter sounds. It consists of two composites,

the Phonological Awareness Composite and Letter Sounds Composite. Only the Phonological Awareness Composite was administered in the current study due to time constraints at the schools that prevented the administration of the Letter Sounds Composite.

Torgesen and Bryant (2004b) provided strong psychometric evidence for the TOPA-2+. The mean internal consistency and test-retest reliability coefficients for the Phonological Awareness Composite at the kindergarten level were .91 and .87, respectively. As evidence of validity, Torgesen and Bryant reported moderate to strong correlations between the TOPA-2+ and a variety of instruments purporting to measure the same or similar constructs. Nelson (in press) found moderate correlations between the TOPA-2+ Phonological Awareness Composite at midkindergarten and the Woodcock–Johnson Tests of Achievement, Third Edition Letter Word Identification and Word Attack subtests at the end of kindergarten, with predictive validity coefficients of .59 and .54, respectively.

The TOPA-2+ was selected as one of the reference standards for the current study because of its strong psychometric properties. Additionally, its use allowed comparison to Hintze et al. (2003) who used a similar norm-referenced measure of phonological awareness. Consistent with Hintze et al., a standard score of lower than 85 was used as the performance standard for distinguishing those with adequate from those with inadequate phonological awareness skills.

Woodcock–Johnson Tests of Achievement, Third Edition

The Woodcock–Johnson Tests of Achievement, Third Edition (WJ III; Woodcock, McGrew, & Mather, 2001) Letter Word Identification (LW), and Word Attack (WA) subtests were administered. The LW and WA subtests make up the WJ III Basic Reading Skills Cluster and measure letter knowledge, letter-sound knowledge, and real and nonsense word reading skill.

McGrew and Woodcock (2001) reported median alpha coefficients of .94 and .87 for the LW and WA subtests, respectively. Furthermore, these authors reported median test-retest reliability coefficients of .95 and .83 for the LW and WA subtests, respectively, when a 1-year interval between testing was used. Finally, as evidence of concurrent validity, McGrew and Woodcock found moderate to strong correlations of the WJ III Basic Reading Skills Cluster with the Kaufman Test of Educational Achievement Reading Decoding scale ($r = .66$) and the Wechsler Individual Achievement Test Basic Reading scale ($r = 82$).

The WJ III was selected as the second of the current study's reference standards because of its excellent psychometric properties. Jenkins (2003) stated, "In the field of early intervention, the family of achievement tests developed by Richard Woodcock and associates come closest to a 'gold standard' criterion test of reading ability" (p. 2). A standard score of 90 was set as the performance standard on the WJ III for distinguishing between adequate and inadequate reading skill because this is a common performance standard used in the reading research (e.g., Siegel, 1989; Speece, Mills, Ritchey, & Hillman, 2003). Additionally, achievement below this cutoff has been recommended as the first requirement in determining whether a child has a reading disability (Dykman & Ackerman, 1992). If participants achieved at or below a standard score of 90 on the WJ III LW subtest, WA subtest, *or* Basic Reading Cluster, they were classified as having inadequate reading skills.

Procedure

The DIBELS and the TOPA-2+ were administered in mid-January. Over a period of 2 days, the TOPA-2+ was administered in small groups of two to five students, a procedure that is permissible according to Torgesen and Bryant (2004b). The DIBELS was then individually administered to participants over 5 days. Each participant was individually administered the WJ III LW and WA subtests in mid-May. Graduate students in school psychology and advanced undergraduate psychology majors received 12 hours of training on the administration of the measures and administered all instruments used in the study. Additional services were not provided based on any of the assessment data.

Data Analyses

Computer software developed by Watkins (2002) was used to calculate a variety of diagnostic accuracy statistics. Test scores were coded in accordance with the pre-established cutoff scores for both the reference standards (1 = reading problem, 0 = no reading problem) and the DIBELS (1 = at risk or some risk, 0 = no risk). Data for each DIBELS task using each reference standard were then entered into a two-by-two contingency table for calculating diagnostic accuracy statistics. The following describes each diagnostic accuracy statistic: (a) sensitivity indicates the percentage of participants with reading problems according to the reference standards who are identified as at risk on the DIBELS, (b) specificity indicates the percentage of participants identified as having adequate reading skills on the reference standards who are identified by the DIBELS as being at low risk, (c) positive predictive power indicates the percentage of participants identified by the DIBELS as at risk who are classified by the reference standards as having inadequate reading skills, (d) negative predictive power indicates the percentage of participants identified by the DIBELS as at low risk who are classified by the reference standards as having adequate reading skills, (e) false positive rate is the percentage of participants who are identified by the DIBELS as being at risk but who are classified as having adequate reading skills on the reference standards, (f) false negative rate is the percentage of participants who are identified as at low risk by the DIBELS but who are classified as having inadequate reading skills on the reference standards, and (g) kappa indicates the level of agreement beyond chance between the DIBELS and the reference standards. Kappas below .40 are generally regarded as poor, whereas those from .40 to .75 and above .75 are considered fair/good to excellent, respectively (Fleiss, 1981). Software developed by Metz (1998) was used to assess the *overall* diagnostic accuracy of each DIBELS task by conducting receiver operating characteristic (ROC) curve analyses. Additionally, this software was used to calculate the area under the curve (AUC) statistics for each DIBELS task and to conduct *z* tests to investigate potential differences between all possible combinations of the tasks. Swets (1996) described AUCs ranging from .50 to .70, .70 to .90, and .90 to 1.0 as indicative of low, medium, and high diagnostic accuracy, respectively.

Results

Table 1 displays the means, standard deviations, and skewness and kurtosis indexes of the assessments. All scores were normally distributed except for the DIBELS NWF task, which was positively skewed and leptokurtic. Mean scores for the TOPA-2+ and WJ III LW subtest were in the average range, whereas those for the WJ III WA subtest and Basic Reading Skills Cluster were above average. A total of 44 participants (24.9% of the sample) scored below the cutoff on the TOPA-2+, whereas 18 participants (10.2% of the sample) achieved below the WJ III cutoff. Table 2 displays the correlations of the DIBELS with the TOPA-2+ and

Table 1
Mean, Standard Deviations, and Skewness and Kurtosis Indexes

Instrument	*M*	*SD*	Skewness	Kurtosis
DIBELS Initial Sound Fluency	18.25	10.13	.42	−.26
DIBELS Letter Naming Fluency	31.85	14.35	.04	−.23
DIBELS Phoneme Segmentation Fluency	13.73	12.21	.65	−.92
DIBELS Nonsense Word Fluency	14.25	12.89	1.4	2.4
TOPA-2+ Phonological Awareness Scale	96.46	14.76	−.33	−.60
WJ III Letter Word Identification Subtest	107.81	11.04	.15	1.05
WJ III Word Attack Subtest	114.80	14.68	−.55	−.17
WJ III Basic Reading Skills Cluster	112.31	13.39	−.17	.05

Note. DIBELS, Dynamic Indicators of Basic Early Literacy Skills; TOPA-2+, Test of Phonological Awareness – Second Edition: Plus; WJ III, Woodcock-Johnson Tests of Achievement, Third Edition; According to http://dibels.uoregon.edu, students in mid-kindergarten who are at low risk should score 25 to 35 on ISF, at least 27 on LNF, at least 18 on PSF, and know some letter sounds on NWF.

Table 2
Correlations of DIBELS With TOPA-2+ and WJ III

	TOPA-2+	WJ III LW	WJ III WA	WJ BRS
DIBELS ISF	.56	.38	.31	.37
DIBELS LNF	.46	.61	.55	.62
DIBELS PSF	.53	.51	.55	.54
DIBELS NWF	.56	.74	.71	.76

Note. All correlations statistically significant ($p < .01$). DIBELS, Dynamic Indicators of Basic Early Literacy Skills; ISF, Initial Sound Fluency; LNF, Letter Naming Fluency; PSF, Phoneme Segmentation Fluency; NWF, Nonsense Word Fluency; TOPA-2+, Test of Phonological Awareness – Second Edition: Plus; WJ III, Woodcock-Johnson Tests of Achievement, Third Edition; LW, Letter Word Identification Subtest; WA, Word Attack Subtest; BRS, Basic Reading Skills Cluster.

WJ III. All correlations were statistically significant. The magnitude of the correlations was generally moderate to strong.

Examination of DIBELS At-Risk and Some Risk Cutoff Scores

Some Risk

Results indicated sensitivity indexes in the 80% to 90% range for the ISF, PSF, and NWF tasks (see Table 3). Sensitivity rates of 53% to 72% were found for the LNF task. With the exception of the LNF task, the DIBELS tasks using the *some risk* cutoff scores showed substantially lower specificity than sensitivity. Specificity and false positive rates sum to 100%. For the ISF, PSF, and NWF tasks, false positive rates ranged from 41% to 73%.

For all DIBELS tasks using both reference standards, substantially higher negative predictive power than positive predictive power was found. Except for the LNF task, negative predictive power of over 90% was found for all DIBELS tasks. Positive predictive power ranged from 13% to 40%. All Kappas were below .40, ranging from .06 to .30.

Table 3
Diagnostic Accuracy Characteristics of DIBELS Using Some Risk and At-Risk Cutoff Scores

	TOPA-2+ PA Composite					WJ III LW, WA, or BRS				
	Sens.	Spec.	PPP	NPP	Kappa	Sens.	Spec.	PPP	NPP	Kappa
ISF										
9[a]	.52	.90	.64	.85	.45	.50	.83	.25	.94	.23
24[b]	.91	.30	.30	.91	.13	.94	.27	.13	.98	.06
LNF										
14[a]	.32	.93	.61	.81	.30	.56	.92	.43	.95	.42
26[b]	.53	.70	.37	.82	.19	.72	.69	.21	.96	.19
PSF										
6[a]	.68	.74	.47	.88	.37	.67	.67	.19	.95	.16
17[b]	.89	.42	.34	.92	.20	.94	.38	.15	.98	.10
NWF										
4[a]	.62	.79	.49	.86	.37	.67	.73	.22	.95	.21
12[b]	.82	.59	.40	.91	.30	.94	.53	.19	.99	.17
All										
Some risk	1.0	.14	.28	1.0	.08	1.0	.12	.11	1.0	.03
At risk	.86	.61	.42	.93	.35	.89	.53	.18	.98	.15

Note. TOPA-2+ PA, Test of Phonological Awareness – Second Edition: Plus Phonological Awareness Composite; WJ III, Woodcock-Johnson Tests of Achievement, Third Edition Letter-Word Identification subtest, Word Attack subtest, or Basic Reading Skills Cluster; ISF, Initial Sound Fluency; LNF, Letter Naming Fluency; PSF, Phoneme Segmentation Fluency; NWF, Nonsense Word Fluency; Sens., Sensitivity; Spec., Specificity; PPP, positive predictive power; NPP, negative predictive power.
[a] At-risk cutoff score.
[b] Some risk cutoff score.

At risk. The PSF and NWF tasks showed higher sensitivity than the ISF and LNF tasks. The ISF and LNF tasks showed sensitivity rates ranging from 32% to 56%. Because sensitivity and false negative rates sum to 100%, these results also indicated false negative rates from 32% to 68% depending on the DIBELS task and reference standard used. Specificity rates ranged from 67% to 93%.

The *at-risk* DIBELS cutoff scores showed negative predictive power above 80% for all tasks. Positive predictive power ranged from 19% to 64%. The ISF task showed agreement beyond random chance with the TOPA-2+, and the LNF task showed agreement beyond random chance with the WJ III. All other Kappas were below .40.

Examination of Alternative Cutoff Scores

75% Sensitivity and Specificity

None of the DIBELS tasks at any of the cutoff scores met the diagnostic accuracy criterion of 75% on both sensitivity and specificity. As indicated in Table 4, cutoff scores of 14 or 15 on the ISF task produced sensitivity and specificity rates closest to 75%. Cutoff scores of 7 to 8 and 5 to 7 approached these diagnostic characteristics for the PSF and NWF tasks, respectively. The LNF task showed the greatest amount of variability depending on the reference standard used.

90% Sensitivity and Highest Specificity

In order to identify 90% of those classified as having inadequate reading skills on the reference standards, the ISF and LNF tasks required cutoff scores of 23 and 43 to 45, respectively. The associated specificity rates of these cutoff scores were low, ranging from 31% to 35% for the ISF task and 20% to 25% for the LNF task. Greater variability depending on the reference standard resulted when exploring a high degree of sensitivity for the PSF and NWF tasks. For both tasks, use of the WJ III as the reference standard resulted in lower cutoff scores to achieve at least 90% sensitivity than when the TOPA-2+ was used.

Table 4
Cutoff Scores Achieving 75% Sensitivity/Specificity and 90% Sensitivity and Highest Specificity

	Sensitivity	Specificity	PPP	NPP	Kappa
ISF					
14[a]	.72	.65	.19	.95	.17
15[b]	.73	.68	.43	.88	.33
23[c]	.94	.27	.13	.98	.06
23[d]	.91	.30	.30	.91	.13
LNF					
25[a]	.72	.72	.22	.96	.23
36[b]	.70	.43	.29	.81	.09
43[c]	.94	.25	.12	.98	.05
45[d]	.91	.20	.27	.87	.06
PSF					
7[a]	.72	.62	.18	.95	.14
8[b]	.70	.62	.38	.86	.25
11[c]	.94	.47	.17	.99	.14
18[d]	.91	.41	.34	.93	.21
NWF					
5[a]	.72	.70	.22	.96	.21
7[b]	.73	.72	.46	.89	.38
9[c]	.94	.60	.21	.99	.22
18[d]	.95	.42	.35	.97	.24

Note. ISF, Initial Sound Fluency; LNF, Letter Naming Fluency; PSF, Phoneme Segmentation Fluency; NWF, Nonsense Word Fluency.
[a] Cutoff score most closely approximating 75% sensitivity and specificity using WJ III as reference standard.
[b] Cutoff score most closely approximating 75% sensitivity and specificity using TOPA-2+ as reference standard.
[c] Cutoff score achieving at least 90% sensitivity and highest specificity using WJ III as reference standard.
[d] Cutoff score achieving at least 90% sensitivity and highest specificity using TOPA-2+ as reference standard.

Examination of Overall Diagnostic Accuracy

All AUC indexes for the DIBELS tasks indicated medium *overall* diagnostic accuracy when the WJ III was used as the reference standard. The AUCs for the ISF, LNF, PSF, and NWF tasks were .71, .81, .74, and .81, respectively. With the TOPA-2+ as the reference standard, the AUCs of ISF (.78), PSF (.77), and NWF (.70) indicated medium diagnostic accuracy, whereas the AUC of .68 of the LNF task indicated low overall diagnostic accuracy. Two-tailed z tests indicated no statistically significant differences in overall diagnostic accuracy between the DIBELS tasks, other than the NWF task showing stronger overall diagnostic accuracy than the LNF task when the TOPA-2+ was used as the reference standard ($z = 2.32$, $p = .02$).

Examination of Performance on All DIBELS Tasks

As displayed in Table 3 (see last two rows), the majority of participants who exhibited inadequate reading skills on the reference standards also performed in the *at-risk* range on at least one of the four DIBELS tasks. Sensitivity rates nearly reached 90%, with associated specificity rates ranging from 53% to 61% depending on the reference standard. High negative predictive power (93% to 98%) was found, but occurred at the expense of substantially lower positive predictive power (18% to 42%). Both Kappas were below .40.

All participants who exhibited inadequate reading skills on the reference standards performed in the *some risk* range on at least one of the four DIBELS tasks. Table 3 shows that these 100% sensitivity rates were associated with specificity rates ranging from 12% to 14%. Likewise, perfect negative predictive power came at the expense of low positive predictive power (11% to 28%). The Kappas were .03 and .08.

Discussion

Results of the current study were generally consistent with those of a similar study by Hintze et al. (2003). In both studies, the DIBELS cutoff scores for *at-risk* status yielded generally low levels of sensitivity. The current study's results indicate that up to 68% of participants, depending on the DIBELS task under consideration, were "missed" by the DIBELS when the *at-risk* cutoff scores were used. This result is highly similar to Hintze et al. who found false negative rates of up to 67% for cutoff scores at or near the *at-risk* range.

Not surprisingly, the more lenient *some risk* cutoff scores allowed for higher sensitivity than the *at-risk* cutoff scores but lower specificity. On all tasks but LNF, sensitivity rates were over 80%. Using the *some risk* cutoff scores, the ISF task had nearly identical diagnostic accuracy characteristics as those found in the Hintze et al. (2003) study, but the LNF task was found to be less sensitive and the PSF task to be more sensitive in the current study. Substantial false positive rates associated with these cutoff scores were also found. False positive rates ranged from a low of 30% to a high of 73%. These false positive rates are similar to those found by Hintze et al., who found false positive rates of up to 61% when using cutoff scores at or near the *some risk* range. Furthermore, the false positive rates found in the current study are consistent with O'Connor and Jenkins (1999) who found false positive rates of 47% to 70% when they adjusted the cutoff scores of their early reading screening measures to achieve high sensitivity.

Although sensitivity and specificity are the most commonly reported diagnostic accuracy statistics, positive and negative predictive power statistics are more clinically useful because they take into account the base rate of the attribute under consideration (Kessel & Zimmerman, 1993). Whereas positive predictive power provides evidence for the use of an instrument as an inclusion criterion, negative predictive power provides evidence for use as an exclusion criterion (McDermott et al., 1995). In both the current study and Hintze et al. (2003), the DIBELS showed greater negative predictive power than positive predictive power. Negative predictive power was over 80% for all DIBELS tasks and often over 90%. None of the DIBELS tasks demonstrated positive predictive power at the level regarded as acceptable for screening measures (.70; Glascoe et al., 1992). Other studies (e.g., O'Connor & Jenkins, 1999; Speece, 2005) have found poor positive predictive power but high negative predictive power for early reading screening instruments. The DIBELS using the *some risk* and *at-risk* cutoff scores appear to possess strong ability to identify those who are adequate readers but less ability to identify those who are inadequate readers.

In addition to examining the diagnostic accuracy of each DIBELS task independently, the current study examined the diagnostic accuracy of the DIBELS tasks when used together. At the *at-risk* level, sensitivity rates of over 85% and moderate specificity rates were found. Again, results indicated high negative predictive power but low positive predictive power. Use of the *some risk* cutoff scores resulted in perfect sensitivity rates. These sensitivity rates of 100% came at the expense of very high false positive rates (86% to 88%). High false positive rates such as these are not unexpected given the large number of students identified as at *some risk* when performance on all the kindergarten tasks was considered together. Nearly 90% ($n = 158$) of participants were classified as at *some risk*, whereas only 10% were classified as at *low risk* ($n = 19$).

Beyond the cutoff scores developed by the authors of the DIBELS, some school professionals may be interested in cutoff scores that produce specific diagnostic accuracy characteristics. Hintze et al. (2003) deemed cutoff scores that produce both sensitivity and specificity rates of at least 75% as adequate. They found that cutoff scores of 15 and 25 on ISF and LNF, respectively, produced diagnostic accuracy characteristics that approximated this criterion, but that no cutoff scores for the PSF task met the criterion. The current study found nearly identical cutoff scores when the WJ III was used as the reference standard.

Those who are highly prevention focused are likely interested in identifying as many children who are at risk as possible so these children can be provided early intervention services. In the current study, cutoff scores that produced sensitivity rates of at least 90% with the highest associated specificity rates were examined. To achieve 90% sensitivity, false positive rates were often substantial, particularly on the ISF and LNF tasks. These tasks showed false positive rates of up to 80% when the cutoff scores were set to achieve 90% sensitivity.

The current study's results provide support for Jenkins' (2003) position that criterion validity only provides weak evidence for the utility of screening measures. Indeed, moderate to strong correlations were found between the DIBELS tasks and the criterion measures. However, the overall diagnostic accuracy characteristics of the DIBELS tasks using a variety of cutoff scores indicated the utility of these measures was moderate. The current study's results indicate that correlational evidence is not enough for determining the utility of early reading screening measures and that the investigation of classification validity is essential.

Because of the challenges of early childhood assessment, it is important to underscore a reasonable level of expectation for reading screening measures. Most would likely agree that it is unrealistic to expect reading screeners to have sensitivity indexes of 100%. All screening measures will produce some degree of error in classification. It is not unrealistic, however, for screeners to have respectable levels of both sensitivity and specificity and positive and negative predictive power. Unquestionably, a trade-off does occur between these diagnostic accuracy properties when various cutoff scores are examined, but *extreme* diminution of one is not requisite for the improvement of the other.

Depending upon the intended use of a screening measure, some diagnostic statistics are weighted more heavily than others. The potential costs and benefits of the decisions being made based on the screening results must be taken into consideration. The position often taken regarding early reading screeners is that their aim should be to identify as many children who are at risk as possible, with more concern given to reducing false negatives than false positives (Felton, 1992). Within an RtI service delivery system, the reduction of false negatives is considered essential because without high sensitivity, those students most in need will not receive the intensive secondary interventions they require to develop adequate reading skills.

Referring to early reading screening measures, Felton (1992) stated, "The problem of false positives is not a major one" (p.7). Unquestionably, the hazards associated with false positives and false negatives vary depending upon the decisions made (Macmann & Barnett, 1999). False positive errors, however, are far from benign in the context of early reading screening. The major negative consequences of making false positive errors include squandering meager instructional resources (Bishop, 2003; Jenkins, 2003; O'Connor & Jenkins, 1999; Speece, 2005) and causing unnecessary parent, teacher, and student anxiety (Swets, Dawes, & Monahan, 2000). Rather than trivializing false positive errors, a balanced perspective in which both types of errors are regarded as undesirable appears more defensible. Allowing an *unreasonable* number of false positive cases in order to identify true positives is questionable (Swets et al., 2000).

Limitations

The current study's results should be interpreted in light of the following limitations. First, generalizability should be considered in relation to the sample's demographic characteristics. Over 90% of participants were White, and all were drawn from the same area of the country. Second, achievement on the WJ III was above average for the current sample. A lower percentage of students were classified as having inadequate reading skills on the WJ III in the current sample than the norm sample of the

instrument. The establishment of strong diagnostic accuracy characteristics becomes more difficult with fewer at-risk cases to predict on the reference standard (Chaffee, Cunningham, Secord-Gilbert, Elbard, & Richards, 1990). A final limitation is inherent in all diagnostic accuracy studies due to their incorporation of a gold standard. The assumption is that this standard is 100% accurate; however, as Macmann and Barnett (1999) stated, "Given the indefinite nature of the constructs measured, 'true' criterion status is known only to God (and even She or He may have questions)" (p. 525). Shortcomings exist for all psychological and educational instruments, even those that might be characterized as gold standards. In the current study, the shortcomings of the WJ III LW and WA subtests, especially for younger children, should be highlighted. According to Bracken and Walker (1997), standard scores on tests should not change by more than one third of a standard deviation with a 1-point raw score change. For kindergarten children, especially those with reading skills near the floor of the instrument, the WJ III violates this principle. Children can achieve large standard score increases by correctly answering only one more item on the WJ III reading tasks. Because the WJ III does not possess an adequate number of items near the floor of the instrument to clearly discriminate between lower levels of reading skill, some participants' scores may have been spuriously inflated. This psychometric shortcoming of the WJ III potentially deflated the diagnostic accuracy characteristics of the DIBELS. It should be noted, however, that the diagnostic accuracy characteristics of the DIBELS were similar even when the TOPA-2+, which does not suffer as severely from the same psychometric limitation, was used as the reference standard.

Implications for Research and Practice

Pertaining to practical implications, the results of both the current study and Hintze et al. (2003) indicate that if the DIBELS is utilized the screening process should not end after its administration. Doing so would likely lead to an excessive number of false positive cases, which doubtless would result in a dilution of instructional services for those who truly need intensive, explicit, and systematic reading instruction. The current study provides strong evidence that the DIBELS has utility as an exclusionary measure due to its adequate negative predictive power, but not an inclusionary measure because of its poor positive predictive power. Perhaps a multistep process could be implemented in which the DIBELS or other universal screener is used to exclude those who are not at risk as a first step. Following universal screening, more specific and thorough assessment could be conducted with those who are not excluded. These assessments would likely need to be individually administered and could include more sophisticated phonological awareness tasks. The assessment of other reading-related skills, such as rapid automatized naming and letter sound knowledge, may also produce more accurate identification of at-risk readers. If upon secondary assessment students are confirmed to be at risk, more intensive secondary interventions could be provided to prevent reading difficulties from developing.

School systems that conduct early reading screening likely place a high premium on identifying all or nearly all children who are at risk for reading failure. That is, they likely desire instruments with high sensitivity. If this is the goal, schools should not use the cutoff scores established for *at-risk* status on individual DIBELS tasks. These cutoff scores produced high false negative rates in the current study. Results of the current study suggest, however, that considering performance on the DIBELS tasks at the *at-risk* level collectively would lead to high sensitivity and moderate specificity. The cutoff scores established for *some risk* status resulted in high sensitivity indexes for all DIBELS tasks except LNF. If performance at the *some risk* level is considered collectively, unreasonably high false positives will likely occur.

Future researchers should continue to search for universal reading screening measures with adequate diagnostic accuracy characteristics. Because kindergarten children are in the acquisition phase of learning to read, finding psychometrically sound instruments to measure "true" reading skill for this age group and thus serve as reference standards continues to be a challenge. Examining the impact of growth on the reading screening process is a potentially fruitful line of research (see Speece, 2005). Due to reading's complex nature and children's variable development, early identification of those at risk for developing reading problems is an inherently

difficult pursuit, but essential in efforts to prevent reading problems.

References

Bishop, A. G. (2003). Prediction of first-grade reading achievement: A comparison of fall and winter kindergarten screenings. *Learning Disability Quarterly, 26,* 189–200.

Bracken, B. A., & Walker, K. C. (1997). The utility of intelligence tests for preschool children. In D. P. Flanagan, J. L. Genshaft, & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (pp. 484–502). New York: Guilford Press.

Chaffee, C. A., Cunningham, C. E., Secord-Gilbert, M., Elbard, H., & Richards, J. (1990). Screening effectiveness of the Minnesota Child Development Inventory Expressive and Receptive Language Scales: Sensitivity, specificity, and predictive value. *Psychological Assessment, 2,* 80–85.

DIBELS Benchmark Goals and Indicators of Risk, Three Assessment Periods per Year. (n.d.). Retrieved August 2, 2007, from http://dibels.uoregon.edu/benchmarkgoals.pdf

Dykman, R. A., & Ackerman, P. T. (1992). Diagnosing dyslexia: IQ regression plus cutpoints. *Journal of Learning Disabilities, 25,* 574–576.

Felton, R. H. (1992). Early identification of children at risk for reading disabilities. *Topics in Early Childhood Special Education, 12,* 212–230.

Fleiss, J. L. (1981). *Statistical methods for rates and proportions* (2nd ed.). New York: Wiley.

Glascoe, F. P., Byrne, K. E., Ashford, L. G., Johnson, K. L., Chang, B., & Strickland, B. (1992). Accuracy of the Denver-II in developmental screening. *Pediatrics, 89,* 1221–1225.

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

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,* 541–556.

Jenkins, J. R. (2003, December). *Candidate measures for screening at-risk students.* Paper presented at the National Research Center on Learning Disabilities Responsiveness-to-Intervention Symposium, Kansas City, MO.

Kame'enui, E. J., Francis, D. J., Fuchs, L., Good, R. H., O'Connor, R. E., Simmons, D. C., et al. (2002). *Analysis of reading assessment instruments for K-3.* Washington, DC: National Institute for Literacy.

Kame'enui, E. J., Fuchs, L., Francis, D. J., Good, R., O'Connor, R. E., Simmons, D. C., et al. (2006). The adequacy of tools for assessing reading competence: A framework and review. *Educational Researcher, 35,* 3–11.

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

Kessel, J. B., & Zimmerman, M. (1993). Reporting errors in studies of diagnostic performance of self-administered questionnaires: Extent of the problem, recommendations for standard presentation of results, and implications for the peer review process. *Psychological Assessment, 5,* 395–399.

Macmann, G. M., & Barnett, D. W. (1999). Diagnostic decision making in school psychology: Understanding and coping with uncertainty. In C. R. Reynolds & T. B. Gutkin (Eds.), *Handbook of School Psychology* (3rd ed., pp. 519–548.). New York: Wiley.

Manzo, K. K. (2005). National clout of DIBELS test draws scrutiny. *Education Week, 25,* 1–2.

McDermott, P. A., Watkins, M. W., Sichel, A. F., Weber, E. M., Keenan, J. T., Holland, A. M., et al. (1995). The accuracy of new national scales for detecting emotional disturbance in children and adolescents. *Journal of Special Education, 29,* 337–354.

McGrew, K. S., & Woodcock, R. W. (2001). *Woodcock-Johnson III technical manual.* Itasca, IL: Riverside.

Metz, C. E. (1998). *Rockit 0.9B.* Chicago, IL: University of Chicago.

Nelson, J. M. (in press). Psychometric evaluation of the Mountain Shadows Phonemic Awareness Scale with a kindergarten sample. *Journal of Psychoeducational Assessment.*

Nelson, J. M., & Machek, G. R. (2007). A survey of training, practice, and competence in reading assessment and intervention. *School Psychology Review, 36,* 311–327.

O'Connor, R. E., & Jenkins, J. R. (1999). Prediction of reading disabilities in kindergarten and first grade. *Scientific Studies of Reading, 3,* 159–197.

Siegel, L. (1989). IQ is irrelevant to the definition of learning disabilities. *Journal of Learning Disabilities, 22,* 469, 478, 486.

Speece, D. L. (2005). Hitting the moving target known as reading development: Some thoughts on screening children for secondary interventions. *Journal of Learning Disabilities, 38,* 487–493.

Speece, D. L., Mills, C., Ritchey, K. D., & Hillman, E. (2003). Initial evidence that letter fluency tasks are valid indicators of early reading skill. *Journal of Special Education, 36,* 223–233.

Swets, J. A. (1996). *Signal detection theory and ROC analysis in psychology and diagnosis: Collected papers.* Mahwah, NJ: Erlbaum.

Swets, J. A., Dawes, R. M., & Monahan, J. (2000). Better decisions through science. *Scientific American, 283,* 82–87.

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

Torgesen, J. K., & Bryant, B. R. (2004a). *Test of Phonological Awareness Plus* (2nd ed.). Austin, TX: Pro-Ed.

Torgesen, J. K., & Bryant, B. R. (2004b). *Test of Phonological Awareness Plus, examiner's manual* (2nd ed.). Austin, TX: Pro-Ed.

Watkins, M. W. (2002). *Diagnostic Utility Statistics* (Computer software). State College, PA: Ed. & Psych Associates.

Woodcock, R. W., McGrew, K. S., & Mather, N. (2001). *Woodcock-Johnson III Tests of Achievement.* Itasca, IL: Riverside.