Center on Teaching & Learning
UO DIBELS Data System

Evaluation of the DIBELS (6th Edition) Diagnostic System for the Selection of English-Proficient Students at Risk of Reading Difficulties

Keith Smolkowski

Oregon Research Institute

Kelli D. Cummings

University of Oregon

Keith Smolkowski, Oregon Research Institute, Eugene, Oregon; Kelli Cummings, College of Education, University of Oregon, Eugene, Oregon.

Please address correspondence to Keith Smolkowski, Oregon Research Institute, 1776 Millrace Drive, Eugene, OR, 97403. E-mail: keiths@ori.org.

Abstract

This manuscript provides a comprehensive evaluation of the Dynamic Indicators of Basic Early Literacy Skills (DIBELS) 6th Edition set of measures and, in the process, gives a practical illustration of signal detection methods, the methods used to determine the value of screening and diagnostic systems. Data were drawn from a sample of 13,507 English-proficient students in kindergarten through Grade 3, with more than 4,500 students per grade level. Results indicate that most DIBELS measures are accurate and that previously published decision thresholds for DIBELS are generally appropriate, with some key exceptions. For example, the performance of phoneme segmentation fluency did not always meet expectations. We focus on the implications of these findings for educators, and make recommendations for applying this methodology to the development and use of diagnostic systems in education.

Keywords: evaluation of diagnostic systems; signal detection theory; Dynamic Indicators of Basic Early Literacy Skills

Evaluation of the DIBELS (6th Edition) Diagnostic System for the Selection of English-Proficient Students at Risk of Reading Difficulties

Over the past decade, screening instruments have risen to prominence in the field of education. No Child Left Behind (NCLB, 2001), the amendment of the Individuals with Disabilities Education Act to include the provision of Response to Intervention (IDEA, 2004), and other educational initiatives (e.g., American Recovery and Reinvestment Act, 2009) have at their core a requirement for universal screening, which typically involves a data-based cut score to identify students as being at risk for poor reading outcomes as well as provisions for early intervention services. The practical benefits of universal screening include a specific focus on prevention and a close connection between student assessment and learning. Some authors (e.g., Stecker, Fuchs, & Fuchs, 2005) have shown that screening systems can help teachers make more efficient and effective decisions, and improve group-level achievement. Others have also demonstrated that screening assessments with empirical cut scores can reduce disproportionality in referrals for special education (Marston, Muyskens, Lau, & Canter).

The purpose of this paper is to apply signal detection methods to select optimal cut scores for Dynamic Indicators of Basic Early Literacy Skills (DIBELS) 6th Edition (Good & Kaminski, 2002) using a large sample of English-proficient students from kindergarten through Grade 3. The complete evaluation of a diagnostic system involves (a) a description of the overall accuracy of the system and (b) definition the specific scores that optimally identify the outcome of interest (e.g., students with reading difficulties). An evaluation of DIBELS 6th Edition measures and their associated decision thresholds is needed. Although schools have begun to adopt DIBELS Next, there has been scant published research on the measures. The data system at the University of Oregon also reports that more than 600,000 students in more than 3,200 schools were assessed with DIBELS 6th Edition in the 2012–2013 school year, and many more students have been assessed with 6th Edition measures by schools that use a different or no data system. Despite the continued, wide use of DIBELS 6th Edition and its highly publicized set of cut points for decision making, no prior peer-reviewed published work has examined both sets of cut points (i.e., at-risk and moderate-risk decisions) for all measures in kindergarten through Grade 3. In addition, DIBELS 6th Edition is well known and highly researched, but the literature does not include much information on the overall screening accuracy or the appropriateness of the recommended cut scores. The results from this study provide accuracy standards that newer screening systems should ideally exceed. By drawing results from 13,507 students proficient in English, with 4,600 to 5,600 in each grade level, our analysis can produce a high degree of precision and removes potential confounds related to English language ability. Thus, this paper documents the first thorough evaluation of DIBELS 6th Edition as a diagnostic system, and, in the process, provides a practical illustration of the methods reviewed in Smolkowski, Cummings, and Strycker (in press)—methods that have not always been effectively applied in the school psychology or education literature.

Prevention-Oriented Universal Screening

Universal screening for risk in the domain of educational achievement was relatively novel at the turn of this century. Until very recently, curriculum-based measures—the tools most commonly associated with present-day academic screening—were nearly exclusive to special education. The more widespread use of screening measures was catalyzed by federal legislation, based in part on empirical research showing that achievement deficits, especially in reading, became increasingly difficult to remedy after Grade 4 (Juel, 1988).

The implementation of universal screening as a recommended practice was codified in a practice guide distributed by the Institute for Education Sciences (Gersten et al., 2008). The guide lists universal screening as the natural first step in any Response to Intervention model, with moderate evidence for its efficacy. Key features of universal screening include (a) assessment of all students at the beginning and middle of each school year, (b) measures that efficiently indicate essential academic skills, and (c) empirically derived cut scores that accurately identify students who are at risk for poor reading outcomes. The guide also recommends effective data displays to help school teams make instructional decisions about groups of students. Although associated with special education eligibility, universal screening is intended to support lower-stakes decisions. The primary aim of using a cut score with universal screening methods is to capture potential deficits as early as possible, and remediate them through careful instructional planning and delivery. Students who score below a cut point on a screening measure may need additional instruction, more frequent monitoring, or both.

Advances in lower-stakes, high-yield screening assessments now make possible “a long-held dream of many [education professionals]” (Gersten & Dimino, 2006, p. 102)—that teachers have ready access to assessment data for guiding instruction (Fuchs & Fuchs, 1986; Gersten, Keating, & Irvin, 1995). Despite these advances, a significant gap remains between what researchers, policy makers, and test developers recommend and what states, districts, and schools are able to implement. In part this gap reflects the many disagreements in the field regarding appropriate cut scores for identifying students who need support.

Evolution of Screeners in Education

Criterion-referenced goal setting. Universal screening is distinct from the practice of collecting local normative information for resource-allocation decisions (e.g., Shinn, 1988). Establishing local norms with curriculum-based measures has been recommended as a way to compare an at-risk student’s performance with his or her most-similar peers (Shinn, 1988, Deno, 2003). Local norms are useful in evaluating students who are underrepresented on standardized, norm-referenced tests (Elliott & L.S. Fuchs, 1997), and can help schools to quickly and objectively identify students at highest risk. Nonetheless, local norms have left some unsatisfied—especially as accountability standards continue to rise—simply because "typical performance does not equal good performance" (Stewart & Silberglitt, 2008, p. 228).

DIBELS. In 2001, Good, Simmons, and Kame'enui published a paper describing the use of reading measures that would: (a) align to the critical areas in beginning reading (also summarized in Kaminski & Good, 1996; 1998) and (b) predict end-of-year state test performance. Their results underscored the importance of screening all students (as in the case of collecting local normative information), and they were the first authors we know of to prescribe the use of predictive cut scores at single points in time as a way to identify students in need of instructional support. The methods in their paper were not new (c.f. Swets, 1973), but led to a sea change in terms of how teachers applied assessment data for screening decisions.

Since the publication of that first paper, the use of DIBELS in schools has evolved much to include participation, by conservative estimates (Cummings, Park, & Bauer Schaper, 2012), of nearly one in six U.S. public schools serving students in kindergarten through Grade 3. Such widespread use warrants an updated investigation of the optimal cut scores for DIBELS 6th Edition—especially given that the basic procedures used to determine the initial decision thresholds in Good et al. (2001) and even a later technical report (Good, Simmons, Kame'enui, Kaminski, & Wallin, 2002) were somewhat ambiguous and determined by multiple criteria.

As an example, in Good et al. (2001) the authors specified that all Grade 1 students should read 40 correct words per minute by the end of the year. This standard served as the anchor for the system. Although the authors stated, "a second criterion of an effective goal is rigor or ambitiousness" (Good et al., 2001, p. 267), it is difficult to determine the specific criteria used to translate this goal into cut scores. As noted in Smolkowski et al. (in press), without knowing the criteria for "healthy" performance, it is impossible to gauge the quality of the screener or its cut points.

Good et al. (2001) and Good, Simmons, et al. (2002) often used a later DIBELS assessment to distinguish the two populations—typically achieving students and students with reading difficulties. This practice violates the recommendation to maintain independence between the diagnostic system under evaluation and the criterion measure (Smolkowski et al., in press). The use of DIBELS as the criterion likely inflates the appearance of accuracy and may bias decision thresholds. Good et al. (2001) and Good, et al. (2002) did not characterize the overall accuracy of measures beyond correlations and most correlations compared different DIBELS measures collected at different times, such as PSF in spring of kindergarten and NWF in the winter of Grade 1. Finally, the evaluation of scatterplots presented in Good et al. (2001) relied on the equivalent of predictive values to confirm the chosen decision thresholds, which, as noted by Smolkowski et al. (in press), depends on the relative number of students in the samples used to represent each of the two populations of interest.

Research Aims

Our evaluation of DIBELS 6th Edition measures is driven by two goals: (a) to evaluate the full set of DIBELS measures from kindergarten through Grade 3 and, in so doing, (b) to provide a practical illustration of the use of signal detection methods for decision making, described in Smolkowski et al. (in press). Based on data collected from Oregon Reading First schools, the present analyses evaluate DIBELS' ability to predict reading problems determined by comprehensive, end-of-year reading assessments in kindergarten and Grades 1, 2, and 3. These analyses extend current research by (a) using comprehensive reading tests as criterion measures collected at the end of each grade level, (b) drawing on a large sample of students from 34 schools across multiple districts, (c) demonstrating the need for methods that are independent from the sample under study, (d) examining only students proficient in English (i.e., those who have not received services for English as a second language), thus removing any potential influence of accuracy and decision threshold stemming from lack of English language proficiency, and (e) providing new recommendations for high risk (20th percentile on the Stanford Achievement Test–10th Edition [SAT–10]; Harcourt Educational Measurement, 2002) and some risk (40th percentile). This is also the first analysis of literacy screeners, to our knowledge, that presents the level of precision along with estimates of accuracy.

Moreover, we introduce the concept of target performance on DIBELS measures. Current recommendations for DIBELS and other measures often suggest cut scores that separate students into high-risk, moderate-risk, and low-risk (benchmark) categories. Benchmark DIBELS goals have been historically associated with performance near the 40th percentile on a comprehensive test. This has led some teachers to accept students who surpass the benchmark decision threshold as reading at a satisfactory level. It is natural to focus on the highest cutoff available (e.g., benchmark)—even if it represents minimally acceptable performance—but some students just above benchmark may not actually achieve the 40th percentile on a high-stakes test such as the SAT–10. A better strategy, introduced here, is to aim for a *target* performance level that corresponds to the 60th percentile on the SAT–10, which represents an ideal reading level, at which students are likely to meet or exceed standards set by their teachers, districts, or states. Such a target should focus teachers on performance that is apt to lead to the long-term success of their students. In sum, this paper evaluates DIBELS data to provide teachers with a target level of performance and to demonstrate how signal detection methods can be used to identify ideal or even superior performance if desired.

Method

Data for this study were collected from students for whom English was their first language in Oregon Reading First schools from 2003–2004 to 2005–2006. Each school administered the SAT–10 at the end of kindergarten and Grades 1 and 2, and the Oregon Assessment of Knowledge and Skills (OAKS; Oregon Department of Education, 2008) section on reading and literature at the end of Grade 3. These criterion data were used to (a) evaluate the overall accuracy of DIBELS measures as classification tests and (b) determine the optimal thresholds or cut points for each.

Participants and Setting

This study included students from 34 Oregon Reading First schools funded in the first cycle of Reading First, described in greater detail in Baker et al. (2008; 2011). The schools represented 16 independent school districts, half in large urban areas and the rest approximately equally divided between mid-size cities and rural areas. In the 2003–2004 school year, 10% of the students received special education services. During the 2002–2003 school year, 69% of students qualified for free or reduced lunch, and more than a quarter (27%) of third graders did not pass minimum proficiency standards on the OAKS. Across Oregon in 2002–2003, 44% qualified for free reduced lunch and 18% of the Grade 3 students did not pass the third-grade OAKS.

About 32% of the students in the sample were identified as English learners. Nearly 80% of the students assessed were sufficiently proficient that they did not qualify for, or receive, services for students with limited English proficiency (LEP). Students who received LEP services will be addressed in a future manuscript. The present sample consisted of 13,507 English-proficient students who provided valid scores on the high-stakes criterion tests. As 6,544 students provided data in multiple years, the sample includes a total of 20,051 criterion test scores: 5,634 in kindergarten, 4,953 in Grade 1, 4,636 in Grade 2, and 4,828 in Grade 3. About half of the students, 49%, were female, and 6.7% were eligible for special education. Students fell into the following racial–ethnic categories: 57% Caucasian, 22% Hispanic or Latino/Latina, 11% African American, 5% American Indian, 4% Asian, and less than 1% Alaskan Native, Hawaiian, Pacific Islander, or "other."

Data were collected during the first three years of Oregon Reading First implementation. The sample comprised students in kindergarten through Grade 3 who had begun kindergarten in 2000–2001 to 2005–2006. In the fall, winter, and spring, students were administered DIBELS measures (Good & Kaminski, 2002) as part of benchmark testing, and virtually all students in kindergarten through Grade 3 participated in three assessments per year. In the spring, students also were administered a high-stakes reading test. A few students, 3% to 4%, were excluded from high-stakes testing because of absences.

Criterion Measures

Stanford Achievement Test–10th Edition. The SAT–10 (Harcourt Educational Measurement, 2002) is a group-administered, norm-referenced test of overall reading proficiency. The measure is not timed, although guidelines with flexible time recommendations are given. Reliability and validity data are strong. Kuder–Richardson reliability coefficients for total reading scores were .97 at Grade 1 and .95 at Grade 2. Correlations between the total reading score and the Otis–Lennon School Ability Test ranged from .61 to .74. We used the 2007 norms based on a representative sample of the U.S. student population, and we used the total reading score as our criterion.

Oregon Assessment of Knowledge and Skills (OAKS). The OAKS, developed by the Oregon Department of Education (ODE, 2008), is an untimed, multiple-choice test administered yearly to all students in Oregon starting in Grade 3. Reading passages representing literary, informative, and practical selections are included in the Grade 3 test. These passages represent selections that students might encounter in school settings and other daily reading activities. Seven individual subtests require students to: (a) understand word meanings in the context of a selection; (b) locate information in common resources; (c) answer literal, inferential, and evaluative comprehension questions; (d) recognize common literary forms, such as novels, short stories, poetry, and folk tales; and (e) analyze the use of literary elements and devices, such as plot, setting, personification, and metaphor. The Oregon Department of Education reports that OAKS criterion validity was .75 with the California Achievement Tests and .78 with the Iowa Tests of Basic Skills (ODE, 2005). The four alternate forms used for the OAKS demonstrated internal consistency reliability (Kuder–Richardson) of .95 (ODE, 2000).

Screener Measures

DIBELS Letter Naming Fluency (LNF). LNF (Kaminski & Good, 2002; Marston & Magnusson, 1988) measures the number of randomly ordered upper- and lowercase letters students name in 1 minute. Alternate-form reliability for this measure was reported as .88 in kindergarten (Good, Wallin, Simmons, Kame‘enui, & Kaminski, 2002), and the concurrent validity of the measure was .70 with the Woodcock–Johnson Psycho-Educational Battery–Revised readiness cluster standard score in kindergarten. The predictive validity was .65 with the Woodcock–Johnson Psycho-Educational Battery–Revised readiness cluster standard score and .71 with Grade 1 oral reading fluency (Good, Wallin, et al., 2002).

DIBELS Phoneme Segmentation Fluency (PSF). PSF (Kaminski & Good, 1996) measures phonemic awareness. Students are scored on the number of correct individual phonemes they segment from words presented verbally by the examiner in 1 minute. The PSF measure has alternate-form reliability of .88 and predictive validity coefficients ranging from .73 to .91 (Kaminski & Good, 1996).

DIBELS Nonsense Word Fluency (NWF). NWF (Kaminski & Good, 1996) measures alphabetic understanding and phonological recoding ability (Cummings, Dewey, Latimer, & Good, 2011; Fien et al, 2008; Good, Baker, & Peyton, 2009). Students are scored on the number of phonemes they correctly identify from consonant–vowel and consonant–vowel–consonant pseudowords (either the sounds of the individual letters or the pseudoword as a unit) in 1 minute. Good and Kaminski (2002) reported alternate-form reliability for NWF ranging from .67 to .87 and concurrent validity coefficients with the readiness subtests of the Woodcock–Johnson Psycho-Educational Battery–Revised ranging from .35 to .55. Fien et al. (2008) reported concurrent validity coefficients of .51 to .76 with the SAT–10 and DIBELS Oral Reading Fluency from kindergarten through Grade 2.

DIBELS Oral Reading Fluency (ORF). DIBELS ORF (Good & Kaminski, 2002) is a 1-minute fluency measure of reading connected text. Students read a set of three passages at the beginning, middle, and end of the year, and their median score at each assessment is recorded. On DIBELS ORF passages, alternate-form reliability drawn from the same level ranged from .89 to .94, and test–retest reliabilities for elementary students ranged from .92 to .97 (Good & Kaminski, 2002). In Oregon specifically, the correlation between DIBELS ORF passages and OAKS Reading administered in Grade 3 was .67 (Good, Gruba, & Kaminski, 2001). Baker et al. (2008) reported concurrent validity of .82 with Grade 1 SAT–10, .80 with Grade 2 SAT–10, and .67 with Grade 3 OAKS.

Data Collection Procedures

ORF measures were administered to students by school-based assessment teams in the fall, winter, and spring. Each assessment team received 1 day of training on DIBELS administration and scoring, followed by additional training from a reading coach at each school who conducted calibration practice sessions with students. To maintain consistency across testers, coaches checked each assessment team member before data collection.

Teachers administered the SAT–10 and the OAKS in the spring. SAT–10 testing was supervised and monitored by Reading First coaches, who were trained by the Oregon Reading

First Center. Coaches provided training on test administration and monitoring to all teaching staff in their building. Coaches used a fidelity implementation checklist to document testing procedures; median fidelity on 18 test-administration items was 98.3%. Grade 3 students were administered the OAKS according to procedures established by the school, district, and state.

Analysis Approach

The analysis of the DIBELS 6th Edition screening system followed the methods outlined in Smolkowski et al. (in press). We first generated ROC curves and calculated the area under curve, *A*, for each measure administered at each time point. Based on the discussion in Swets (1988), an excellent screener would produce values of *A* above .95, for a very good screener *A* should ranges from .85 to .95, and reasonable screeners yield moderate *A* values from .75 to .85. Values below .75 represent relatively poor diagnostic utility. In reading instruction, we believe teachers can judge student performance reasonably well (Martin & Shapiro, 2011), and their judgments are likely more valuable than the results from a weak screener where $A < .75$.

The decision rule for selecting a threshold (cut score) for each level of risk should depend on the anticipated consequences. Research has linked poor reading achievement to a number of negative outcomes, such as lower socioeconomic status (e.g., Ritchie & Bates, 2013), but the literature does not offer useful approximations of the actual costs and benefits associated with the four potential outcomes (false and true positives and negatives). In such situations, Swets, Dawes, and Monahan (2000) suggests setting decision thresholds based on sensitivity or specificity. We believe it more ethical to provide additional instruction that a student may not technically require than the failure to offer such instruction to a student who truly needs help. If a typically achieving student performs poorly on a screening assessment—a false positive—he or she may be incorrectly assigned to small-group instruction. His or her teacher might quickly determine that the small-group participation is unnecessary and place that student back into standard instruction. Such quick identification and correction of false negatives is far less likely. Due to increased attention from teachers for false positives, those screening errors are relatively easy for teachers to identify. Students with false-negative results, in contrast, are unlikely to receive additional support or progress monitoring because the screener inaccurately classified them as typically achieving students. Students assumed to be typically achieving may not be tested again until the next screener administration for all students, which may inadvertently delay the needed instructional supports. We therefore chose decision thresholds based on the assumption that false-negative errors are more costly than false-positive errors in most instructional settings.

To translate this logic into a decision rule for setting decision thresholds, we assumed that we would want no more than 20% false negatives, which is equivalent to a true positive fraction (TPF, sensitivity) of .80. Hence, for optimal decision thresholds we chose screener scores associated with sensitivity at or above .80. This decision rule also hinges partially on the fact that most reading screeners are not highly accurate (i.e., $A < .95$), so specificity seldom exceeds .80 for the decision thresholds chosen for sensitivity value .80. Finally, this approach to decision thresholds allows for a consistent interpretation of the cut scores for all measures at all administrations, unlike approaches that are ambiguous (e.g., Good et al., 2001) or that depend on a combination of sensitivity and specificity (Silberglitt & Hintze, 2005, Youden, 1950).

For each measure, we reported *A*, the decision threshold, sensitivity, specificity, negative and positive predictive values, and the proportion of students who screened positive (τ). We also describe the level of precision surrounding estimates of *A*, sensitivity, and specificity. These statistics were defined for students at risk (20th normative percentile), students at benchmark (40th normative percentile), and students at the target score (60th normative percentile). The percentiles represent the base rates, ρ. All analyses were conducted with SAS (SAS Institute,

2009) PROC LOGISTIC to estimate *A* and PROC FREQ for other statistics. For reporting, we followed the STARD Statement (STAndards for the Reporting of Diagnostic accuracy studies; http://www.stard-statement.org/) recommendations.

Results

Table 1 provides descriptive information for the SAT–10, OAKS, and DIBELS measures. Tables 2 through 5 present the accuracy, decision thresholds, and related statistics for LNF, PSF, NWF, and ORF at the three levels of performance on the criterion measures. Each decision threshold was chosen as the minimum score that maintained a sensitivity level above .80, so that at least 80% of students in the reading-difficulty population would be identified. Results for LNF are explained below (and in Table 2); the remaining results may be interpreted similarly.

Results Example: LNF

In Table 2, we report the results for LNF. For the at-risk level, the accuracy for DIBELS LNF in the fall of kindergarten was relatively low, *A* = .77 with 95% confidence interval (CI) of [.76, .78]. The area under the ROC curve is just above.75, the value chosen as minimally acceptable. Students who were truly at risk of reading failure (sensitivity) had an 81% chance of being identified as such if they scored below 6 on LNF, 95% CI = [.80, .82]. Specificity values were less than ideal. Of the students not below the 20th percentile, 62% were likely to be identified as a true negative, 95% CI [.61, .63]. This translates into a 38% chance that students without risk of falling below the 20th percentile on the SAT–10 would be falsely identified as a positive, which may be too high in some schools. A different threshold for LNF could improve specificity but only at the expense of reduced sensitivity. The challenge with LNF collected in the fall of kindergarten does not lie with the decision threshold but rather with the low accuracy of LNF at this time period (*A* = .77). The winter administration in kindergarten had a higher level of accuracy, *A* = .84, 95% CI [.83, .85], and consequently a more-acceptable specificity value, .71, 95% CI [.70, .72], for the chosen level of sensitivity. Nonetheless, the overall accuracy of LNF rarely exceeded the moderate range, *A* = .75 to .85.

Predictive values suggest the "clinical" value of the screener (Pepe, 2003). Among the 56% of students (τ) who scored below 6 on the fall assessment of LNF in kindergarten and thus screened positive, the positive predictive value (PPV) shows that 62% will likely fall below the 20th percentile on the SAT–10. For the same assessment, 81% of students who screened negative will likely score at or above the 20th percentile on the SAT–10 in the spring. For a school with a similar base rate, the PPV and negative predictive value (NPV) quantify the clinical implications of leaving students unsupported. PPV and NPV, however, depend on the base rate (ρ). PPV can range only between ρ and 1, and NPV is restricted to the range between ρ – 1 and 1, so for the at-risk threshold in the fall of kindergarten PPV must lie between .43 and 1 and NPV values between .67 and 1. Hence, NPV will always be fairly high for risk levels with a low base rate. In contrast, the PPV will be high for criteria with high base rates, such as with the target criterion with $\rho = .82$.

But because most schools have different base rates than those reported here, the predictive values in Table 2 will not likely generalize. It is possible to recalculate predictive values and τ, the proportion screened positive, for a different base rate value of ρ and the TPF and false positive fraction (FPF; 1 – specificity) values from the tables as follows (Pepe, 2003):

$$\text{PPV} = \rho \cdot \text{sensitivity} / (\rho \cdot \text{sensitivity} + (1 - \rho) \cdot (1 - \text{specificity}))$$

$$\text{NPV} = ((1 - \rho) \cdot \text{specificity}) / ((1 - \rho) \cdot \text{specificity} + \rho \cdot (1 - \text{sensitivity}))$$

$$\tau = \rho \cdot \text{sensitivity} + (1 - \rho) \cdot (1 - \text{specificity})$$

For the winter LNF assessment in kindergarten, the PPV for the at-risk threshold was .62 given a base rate of .43. In a school with just 15% of its students below the 20th percentile, using the formulas above, the PPV decreases to .27, the NPV increases to .95, and τ changes to .44.

Overall Accuracy of DIBELS

All administrations of LNF, NWF, and ORF demonstrated adequate accuracy ($A > .75$). NWF and ORF were particularly accurate, with some administrations of ORF reaching accuracy levels of $A \geq .90$. In contrast, only the winter kindergarten administration of PSF achieved $A \geq .75$, with some administrations barely surpassing chance, such as the spring of Grade 1, where $A = .56$. Confidence bounds around accuracy values provide important information about precision and uncertainty. Because of the large sample size, confidence intervals for all *A* values fell within ± .02; we reported details in the notes to Tables 2 to 5. Figure 1 displays three representative ROC curves, with their precision, for each measure.

Decision Thresholds

The decision threshold represents the score at which students were no longer identified as a member of the reading-difficulty population for a given level of risk defined by the SAT–10 percentile. Decision thresholds were chosen for the lowest score where sensitivity exceeded .80. Students who scored below 27 on LNF in the winter of kindergarten, for example, were identified as members of the at-risk reading-difficulty population (SAT–10 < 20th percentile). This cut score is much higher than the value selected by Good et al. (2002), who recommended a score below 27 to indicate some risk. The present study suggests that LNF < 34 discriminates some risk from low risk. Tables 2 to 5 list the chosen decision thresholds for each measure and administration. The 95% confidence intervals for sensitivity and specificity were less than ± .01 for nearly all decision thresholds. Figure 1 shows the 95% confidence interval for sensitivity and specificity as a small box around the chosen decision threshold for each ROC curve.

Figures 2 and 3 depict the NWF and ORF decision thresholds by grade level and administration time. These figures suggest a number of insights. The scores that indicate at-risk, some-risk, and target levels of performance increased substantially across each school year and dropped during the summer breaks. Also, although the present analyses produced decision thresholds generally similar to some of the original scores recommended by Good et al. (2001) and Good, Simmons, et al. (2002), others differed markedly. For example, the original thresholds to identify students at risk with NWF (dashed lines in Figure 2) underestimated the performance that students require to reach the 20th percentile on the SAT–10. Finally, some of the original cut scores recommended by Good, Simmons, et al. (2002) were maintained at a constant level, such as for NWF in the winter and spring of Grade 1 and fall of Grade 2. In contrast, results of the current analysis suggest that teachers should focus on improving student performance throughout Grade 1.

Discussion

The present paper applied signal detection methods to DIBELS 6th Edition as a diagnostic system for students from kindergarten through Grade 3. The ability of DIBELS to identify unsuccessful students was tested within an effective, research-based, tiered model of reading instruction provided by Oregon Reading First (Baker et al., 2011). Results demonstrated that most DIBELS measures were accurate with a notable exception: for PSF, the area under the ROC curve was insufficient to recommend that teachers base decisions on this measure after the winter of kindergarten. Low correlations between PSF and later measures have been reported previously (e.g., Good et al., 2001), but such reports had not included diagnostic accuracy. The accuracy of the three other measures—LNF, NWF, and ORF—indicates that they can likely improve decision making.

The decision thresholds chosen for each of the DIBELS measures in this study provide optimal cut scores based on the known likelihood of known outcomes. Specifically, for the two risk classifications, the decision thresholds were chosen to (a) accurately identify at least 80% of students who fall below the 20th percentile on the SAT–10 administered at the end of the school year as having substantial risk of failure and (b) accurately identify at least 80% students who fall below the 40th percentile on the SAT–10 administered at the end of the school year with reduced, but nonetheless some, risk of failure. Students in the first category likely require additional individual or small-group supports, and those in the second category may require additional supports or monitoring to ensure adequate progress throughout the school year. These decision thresholds improve the likelihood that students with similar deficiencies in reading skills will be treated consistently across classrooms, schools, and districts.

The decision thresholds reported here differ somewhat from those previously recommended (Good et al., 2001; Good, Simmons, et al. 2002). The differences, however, are not unexpected, as the previous approach to recommended cut scores employed different methods and a different criterion. The results in Good et al. (2001) also relied on fewer cases—302 to 378, depending on the grade level. A later technical report (Good, Simmons, et al., 2002) drew on a much larger sample, but the report provides mostly descriptive information about the performance of students in the various risk categories. The present analysis was based on a large sample, and identified a decision threshold only for students proficient in English.

Even with substantial differences in methods, decision thresholds reported here were similar to previous findings for some administrations. The present analysis, however, produced LNF cut scores considerably higher than past recommendations, especially after the fall kindergarten assessment. Figures 2 and 3 show the differences in cut scores for NWF and ORF. For NWF (Figure 4), the original cut scores remained constant after the winter of first grade. Findings here demonstrate that gains in predictive utility are available in the winter and spring of Grade 1, especially for students at risk of reading difficulties. Finally, for ORF, the present results generally agree with those in Good et al. (2001) and Good, Simmons, et al. (2002), except when predicting at-risk students in Grade 1 and Grade 3. Differences between the decision thresholds in current use (Good et al., 2001; Good, Simmons, et al., 2002) and those reported here indicate that an update in screener performance standards may improve the efficient delivery of supports. The past cut scores, especially for those markedly below the present decision thresholds, may also offer false hope to students and teachers.

The purpose of this paper was to select optimal cut scores for DIBELS 6th Edition using a large sample of students from kindergarten through Grade 3. Our results provide additional evidence of DIBELS 6th Edition as a valid and accurate diagnostic system supported by substantial research. This comprehensive analysis also offers a standard with which new measures can be compared. Although DIBELS 6th measures were accurate and offer strong diagnostic utility, there is room for improvement. Unfortunately, however, alternative measures do not have published reports documenting their diagnostic accuracy and optimal cut scores. For example, while schools have begun to adopt DIBELS Next, there these measures have little published research support to date and no peer-reviewed research on its diagnostic utility. We therefor recommend that schools continue to use DIBELS 6th Edition, with updated decision thresholds recommended herein, until researchers have had the opportunity to demonstrate that competing measures exceed the accuracy and can improve decision making over DIBELS 6th.

Target Performance

The present investigation introduces a new concept: target performance. The target threshold is meant to help teachers focus on ideal performance. Anecdotal evidence from educators and professional coaches suggests that many teachers work to ensure that students meet the “benchmark” level of performance, which generally corresponds to the 40th percentile on the SAT–10 or other criterion tests. For the average student, however, this level is too low. For students who regularly meet standards, teachers often want to inspire higher performance on end-of-year comprehensive tests. The target performance threshold allows teachers to encourag their students to achieve a higher goal. It also offers teachers an indication of how much better students must perform to move away from the benchmark/some-risk boundary.

In addition to the benefits in the classroom, the target performance decision threshold demonstrates that the methods for the evaluation of screeners can be used to discriminate between populations of high achievers and typically achieving students. The process for assigning target threshold values was identical to those for at-risk and benchmark cut scores except that the criterion was changed to the 60th percentile on the SAT–10.

Limitations

The data for this study were generated from English-proficient students attending schools that participated in Oregon Reading First (Baker et al., 2011). Decision thresholds presented here may not generalize to all children in all schools. The use of sensitivity to set decision thresholds, which is not sensitive to base rates, unlike predictive values, should minimize differences across any schools that aim to achieve the same criterion level of performance.

The quality of the criterion can influence the results of any study of diagnostic systems (Pepe, 2003). Because an imperfect criterion may depress accuracy, the accuracy of most DIBELS measures would likely remain acceptable with a different criterion, but the specific decision thresholds could change.

Except for the measures collected in the fall of kindergarten, the decision threshold for each measure assumes that students received instruction aligned with the content of the measure, as was the case in Oregon Reading First schools. In some instances, the content of the measures may not align with the scope and sequence of the curriculum used in the schools, which could influence the validity of the decision thresholds. For instance, while the present study indicates that a PSF benchmark threshold of 25 indicates some risk for students in the winter of kindergarten, this determination of risk may not be valid for students who receive instruction with curricula that do not introduce the phoneme segmentation skills assessed by DIBELS until after the winter assessment (e.g., Read Well Kindergarten; Sprick, Jones, Dunn, & Gunn, 2008). Thus, below-benchmark performance on PSF may not reflect insufficiencies with either the PSF measure or the instruction but may instead indicate that students had simply not yet been taught the skills that PSF measures. Curriculum-dependent variations are relatively rare, yet may occur for specific measures at certain times, such as for PSF and possibly NWF in the winter of kindergarten. This interaction between instruction and a decision threshold can be critical for affected teachers and requires further investigation. The choice to ignore a decision threshold should be justified, however, through careful inspection of the scope and sequence of curricula.

Future Directions

Numerous other screening systems exist for literacy, and more for numeracy, student behavior, and other domains. Few have been thoroughly evaluated, and many published studies use small samples that result in limited precision (i.e., large confidence bounds; e.g., Hintz, Ryan, & Stoner, 2003). For example, Nelson (2008) examined the classification validity of DIBELS with just 177 kindergarten students but offer no estimates of precision. A quick examination the paper shows that for some estimates, the sample was too small to calculate confidence bounds, and for other estimates, the confidence intervals were wide. In one test, NWF produced a sensitivity value of .62, but using a normal approximation formula (Harper & Reeves, 1999) produced a confidence interval of ± .14 or [.48, .76], which covers the range from very poor to moderate sensitivity values. The present study is one of the first to specify the precision of the estimates, and the broad confidence bounds for the estimates presented in Nelson (2008) demonstrate why reporting precision is important.

The present evaluation did not explore the use of combined or chained screeners. Researchers may combine multiple screeners and choose a decision threshold based on the combination score. In areas of literacy and numeracy, many of the available measures assess specific skills and identify important skill deficits (e.g., decoding versus phonemic awareness; number versus special operations), suggesting that a combination of measures may not yield the most appropriate screening tool. Evaluations of combinations of diagnostic tests may be conducted with Believe the Positive and Believe the Negative rules as well as more sophisticated methods (Pepe, 2003). Gigerenzer and Goldstein (1996), however, suggest the Take The Best heuristic. Consistent with their work, McGrath (2008) showed that “the best single predictor often can perform better than do multiple predictors when the predictors are combined using methods common in applied settings” (p. 195). The costs and benefits of combined or chained screeners and screener/comprehensive-test combinations have yet to be evaluated in education.

Implications for Practice

This evaluation demonstrated that DIBELS 6th Edition measures are generally accurate and that specific, meaningful decision thresholds can be chosen to identify students who require supports. Because the decision thresholds presented here were rigorously evaluated, schools may choose to update their standards based on these findings, and possibly reduce the use of PSF beyond kindergarten. The results of this study do not, however, prescribe the supports required by struggling students. Some students may need only additional data collection, such as progress monitoring or more tests, which may suggest falsely identified as a struggling reader. Others may benefit from intensive supports as soon as their teachers can provide them. This evaluation of DIBELS 6th Edition is silent on the choice and level of supports provided to children.

Many other screening systems exist, with others sure to arrive soon, but new does not guarantee better. Like instructional fads (Slavin, 1989), schools, districts, and states are quick to adopt the next assessment system. Many schools, for example, have begun to adopt DIBELS Next. We caution schools against adopting systems until they have been thoroughly evaluated.

While the properties of most other screening systems, including those used for other academic subjects and student behavior, currently lack supporting research, DIBELS 6th Edition can continue to serve students well.

References

American Recovery and Reinvestment Act [ARRA]. (2009). Public law 111-5 and Public Law 111-8. Division A., Title XIV. State Fiscal Stabilization Fund.

Baker, S. K., Smolkowski, K., Katz, R., Fien, H., Seeley, J. R., Kame‘enui, E. J., & Thomas Beck, C. (2008). Reading fluency as a predictor of reading proficiency in low performing high poverty schools. *School Psychology Review*, *37*(1), 18–37.

Baker, S. K., Smolkowski, K., Smith, J. M., Fien, H., Kame‘enui, E. J., & Thomas Beck, C. (2011). The impact of Oregon Reading First on student reading outcomes. *Elementary School Journal*, *112*(2), 307–331. doi: 0013-5984/2011/11202-0005

Cummings, K.D., Dewey, B., Latimer, R., & Good, R.H. (2011). Pathways to word reading and decoding: the roles of automaticity and accuracy. *School Psychology Review*, *40*(2), 284-295.

Cummings, K.D., Park, Y., & Bauer Schaper, H.A. (2012). Form effects on DIBELS Next oral reading fluency progress-monitoring passages. *Assessment for Effective Intervention*, *38*(2), 91-104.

Deno, S. L. (2003). Curriculum-based measures: Development and perspectives. *Assessment for Effective Intervention*, *28* (3-4), 3-12

Elliott, S.N., & Fuchs, L.S. (1997). The utility of CBM and performance assessment as alternatives to traditional intelligence tests. *School Psychology Review, 26,* 224 – 233.

Fien, H., Baker, S. K., Smolkowski, K., Smith, J. M., Kame‘enui, E. J., & Thomas Beck, C. (2008). Using nonsense word fluency to predict reading proficiency in K–2 for English learners and native English speakers. *School Psychology Review*, *37*(3), 391–408.

Fuchs, L. S., & Fuchs, D. (1986). Effects of systematic formative evaluation: A meta-analysis. *Exceptional Children, 53*(3), 199–208.

Gersten, R., Compton, D., Connor, C. M., Dimino, J., Santoro, L., Linan-Thompson, S., & Tilly, W. D. (2008). Assisting students struggling with reading: Response to Intervention and multi-tier intervention for reading in the primary grades. A practice guide. (NCEE #2009-4045).

Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education. . Retrieved from http://ies.ed.gov/ncee and http://ies.ed.gov/ncee/wwc/publications/practiceguides/

Gersten, R., & Dimino, J. A. (2006). RTI (Response to Intervention): Rethinking special education for students with reading difficulties (yet again). *Reading Research Quarterly*, *41*(1), 99-108.

Gersten, R., Keating, T. J., & Irvin, L. K. (1995). The burden of proof: Validity as improvement of instructional practice. *Exceptional Children*, *61*, 510–519.

Gigerenzer, G., & Goldstein, D. (1996). Reasoning the fast and frugal way: Models of bounded rationality. *Psychological Review*, *103*, 650–669.

Good, R. H., III, Baker, S., & Peyton, J. (2009). Making sense of nonsense word fluency: Determining adequate progress in early first grade reading. *Reading and Writing Quarterly*, *25*, 33–56.

Good, R. H., Gruba, J., & Kaminski, R. (2001). Best practices in using Dynamic Indicators of Basic Early Literacy Skills (DIBELS) in an outcomes-driven model. In A. Thomas & J. Grimes (Eds.), *Best practices in school psychology* (vol. 4, pp. 679–700). Washington, DC: National Association of School Psychologists.

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

Good, R. H., III, Simmons, D. C., & Kame‘enui, E. J. (2001). The importance and decision-making utility of a continuum of fluency-based indicators of foundational reading skills for third-grade high-stakes outcomes. *Scientific Studies of Reading*, *5*(3), 257–288.

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

Good, R. H., III, Wallin, J., Simmons, D. C., Kame‘enui, E. J., & Kaminski, R. A. (2002). *System-wide percentile ranks for DIBELS benchmark assessment* (Technical Report No. 9). Eugene, OR: University of Oregon, Center for Teaching and Learning.

Harcourt Educational Measurement. (2002). *Stanford Achievement Test* [SAT–10]. San Antonio, TX: Author.

Harper, R., & Reeves, B. (1999). Reporting of precision of estimates for diagnostic accuracy: A review. *British Medical Journal*, *318*, 1322-1323.

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*(4), 541-556.

Individuals with Disabilities Education Improvement Act [IDEA 2002]. (2004). Public Law 108-446(20 U.S.C. 1400 *et seq*.).

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

Kaminski, R., & Good, R. (1996). Toward a technology for assessing basic early literacy skills. *School Psychology Review*, *25*(2), 215–227.

Kaminski, R. A., & Good, R. H. (1998). Assessing early literacy skills in a problem solving model: Dynamic Indicators of Basic Early Literacy Skills. In M. R. Shinn (Ed.),*Advanced applications of Curriculum-Based Measurement* (pp. 113-142). New York: Guilford.

Marston, D., & Magnusson, D. (1988). Curriculum-based measurement: District level implementation. In J. Graden, J. Zins, & M. Curtis (Eds.), *Alternative educational delivery systems: Enhancing instructional options for all students* (pp. 137-172). Washington, DC: National Association of School Psychology.

Marston, D., Muyskens, P., Lau, M., & Canter, A. (2003). Problem-solving model for decision making with high-incidence disabilities: The minneapolis experience. *Learning Disabilities Research and Practice, 18*(3), 187-200. doi: 10.1111/1540-5826.00074

Martin, S. D., & Shapiro, E. S. (2011). Examining the accuracy of teachers' judgments of DIBELS performance. *Psychology in the Schools*, *48*(4), 343–356.

McGrath, R. E. (2008). Predictor combination in binary decision-making situations. *Psychological Assessment*, *20*(3), 195–205. doi:10.1037/a0013175

Nelson, J. M. (2008). Beyond correlational analysis of the Dynamic Indicators of Basic Early Literacy Skills (DIBELS): A classification validity study. *School Psychology Quarterly*, *23*(4), 542-552.

No Child Left Behind Act [NCLB]. (2001). Public Law 107-15.

Oregon Department of Education. (2008). *OAKS—test administration manual: 2008–2009 school year*. Retrieved from http://www.ode.state.or.us/teachlearn/testing/manuals/2009/0809tam.pdf

Pepe, M. S. (2003). The statistical evaluation of medical tests for classification and prediction. New York: Oxford.

Ritchie, S. J., & Bates, T. C. (2013). Enduring links from childhood mathematics and reading achievement to adult socioeconomic status. *Psychological Science, 24*(7), 1301-1308. doi: 10.1177/0956797612466268

SAS Institute. (2009). *Base SAS® 9.2 procedures guide: Statistical procedures* (2nd ed.). Cary, NC: SAS Institute, Inc. Retrieved March 10, 2010, from the SAS Product Documentation web site: http://support.sas.com/documentation/index.html

Shinn, M.R. (1988). Development of curriculum-based local norms for use in special education decision making. *School Psychology Review*, *17*(1), 61 – 80.

Silberglitt, B., & Hintze, J. (2005). Formative assessment using CBM–R cut scores to track progress toward success on state-mandated achievement tests: A comparison of methods. *Journal of Psychoeducational Assessment*, *23*, 304–325.

Slavin, R. E. (1989). PET and the pendulum: Faddism in education and how to stop it. *Phi Delta Kappan*, *70*(10), 752-758.

Smolkowski, K., Cummings, K., & Strycker, L. (in press). Fluency measures, their evaluation, and the methods to select students at risk for reading difficulties illustrated with DIBELS (6th Edition) measures. In Cummings, K. D., & Petscher, Y. (Eds.), *The fluency construct*. New York: Springer.

Sprick, M., Jones, S. V., Dunn, R., & Gunn, B. (2008). *Read Well Kindergarten: Critical Foundations in Beginning Reading, 2nd Edition*. Longmont, CO: Sopris West.

Stecker, P. M., Fuchs, L. S., & Fuchs, D. (2005). Using curriculum-based measurement to improve student achievement: Review of research. *Psychology In The Schools*, *42*(8), 795-819. doi:10.1002/pits.20113

Stewart, L. H., & Silberglitt, B. (2008). Best practices in developing academic local norms. In A. Thomas and J. Grimes (Eds.). *Best Practices in School Psychology V* (pp. 225-242). Bethesda, MD: National Association of School Psychologists.

Swets, J. A. (1973). The relative operating characteristic in Psychology. *Science*, *182*, 990–1000.

Swets, J. A. (1988). Measuring the accuracy of diagnostic systems. *Science*, *240*(4857), 1285-1293.

Swets, J. A., Dawes, R. M., & Monahan, J. (2000). Psychological science can improve diagnostic decisions. *Psychological Science in the Public Interest*, 1(1), 1–26.

University of Oregon, Center on Teaching and Learning (2012a). *2012–2013 DIBELS Data System update Part I: DIBELS Next composite score* (Technical Brief No. 1202). Eugene, OR: Author. Retrieved from https://dibels.uoregon.edu/docs/techreports/DDS2012TechnicalBriefPart1.pdf

University of Oregon, Center on Teaching and Learning (2012b). *2012–2013 DIBELS Data System update Part II: DIBELS Next benchmark goals* (Technical Brief No. 1203). Eugene, OR: Author. Retrieved from https://dibels.uoregon.edu/docs/techreports/DDS2012TechnicalBriefPart2.pdf

University of Oregon, Center on Teaching and Learning (2012c). *DIBELS Next recommended benchmark goals: Technical supplement* (Technical Report 1204). Eugene, OR: Author.

Youden, W. J. (1950). Index for rating diagnostic tests. *Cancer*, *3*, 32–35.

Table 1

Descriptive Information for the DIBELS, SAT–10, and OAKS Measures

		Fall			Winter			Spring		
		N	M	SD	N	M	SD	N	M	SD
K	LNF	4981	9.2	12.2	5367	26.6	17.3	5594	39.6	18.1
	PSF				5366	23.4	16.0	5594	45.9	16.0
	NWF				5360	16.7	16.4	5595	34.0	20.7
	SAT–10 percentile							5634	31.2	26.1
Grade 1	LNF	4387	32.9	17.5						
	PSF	4387	31.3	17.7	4701	47.8	15.2	4883	51.8	12.3
	NWF	4387	24.5	21.9	4702	52.4	27.5	4885	70.9	33.3
	ORF				4701	27.8	29.6	4885	51.0	34.9
	SAT–10 percentile							4953	35.5	27.9
Grade 2	NWF	4078	57.7	32.7						
	ORF	4138	41.7	31.8	4408	69.4	39.1	4571	86.4	39.3
	SAT–10 percentile							4635	35.5	27.3
Grade 3	ORF	4393	67.4	35.5	4637	85.8	38.7	4734	103.1	37.7
	OAKS raw score							4828	210.7	11.2

Note. LNF = letter naming fluency; PSF = phoneme segmentation fluency; NWF = nonsense word fluency; ORF = oral reading fluency. The table reports percentiles for the SAT–10, as raw and standard scores are less intuitive. For the OAKS, raw scores are reported; percentile scores were unavailable.

Table 2

Optimal LNF Cut Scores for English-Language-Proficient Students

		Kindergarten			1st
	Statistic	F	W	S	F
At Risk	*A*	.77	.84	.84	.82
	Threshold	**6**	**27**	**42**	**33**
	Sensitivity	.81	.81	.81	.82
	Specificity	.62	.71	.68	.65
	NPV	.81	.82	.82	.87
	PPV	.62	.69	.67	.56
	ρ	.43	.44	.45	.35
	τ	.56	.51	.54	.51
Some Risk	*A*	.79	.85	.85	.82
	Threshold	**11**	**34**	**47**	**38**
	Sensitivity	.81	.81	.80	.81
	Specificity	.63	.71	.72	.64
	NPV	.59	.61	.60	.71
	PPV	.83	.87	.87	.75
	ρ	.70	.70	.71	.58
	τ	.68	.65	.65	.62
Target	*A*	.82	.88	.86	.80
	Threshold	**14**	**37**	**50**	**42**
	Sensitivity	.81	.81	.81	.81
	Specificity	.67	.77	.76	.60
	NPV	.44	.46	.45	.53
	PPV	.92	.94	.94	.86
	ρ	.82	.82	.83	.74
	τ	.73	.70	.71	.71

Note: Letter Naming Fluency (LNF) cut scores were based on SAT–10 or OAKS criterion values at the 60th percentile for target, 40th percentile for some risk, and 20th percentile for at-risk. *A* represents the area under the ROC curve; NPV is the negative predictive value; PPV is the positive predictive value; ρ is the base rate; and τ is the proportion screened positive (scored below the decision threshold). Thresholds bolded if *A* ≥ .75. The 95% confidence intervals were ± .01 for all *A* values, except for target in the fall of kindergarten (± .02), and all sensitivity and specificity values.

Table 3

Optimal PSF Cut Scores for English-Language-Proficient Students

		Kindergarten		1st		
	Statistic	W	S	F	W	S
At Risk	*A*	.79	.73	.73	.68	.60
	Threshold	**28**	54	40	56	61
	Sensitivity	.81	.80	.81	.80	.82
	Specificity	.60	.46	.47	.40	.27
	NPV	.80	.74	.82	.78	.72
	PPV	.61	.55	.45	.43	.39
	ρ	.44	.45	.35	.36	.37
	τ	.58	.65	.63	.68	.77
Some Risk	*A*	.79	.71	.71	.64	.56
	Threshold	**33**	57	44	59	62
	Sensitivity	.80	.81	.81	.81	.82
	Specificity	.62	.40	.39	.32	.23
	NPV	.57	.47	.60	.54	.46
	PPV	.83	.76	.64	.63	.61
	ρ	.70	.71	.58	.59	.59
	τ	.68	.75	.73	.76	.80
Target	*A*	.80	.70	.69	.63	.56
	Threshold	**36**	58	45	60	62
	Sensitivity	.82	.81	.80	.82	.81
	Specificity	.62	.39	.39	.30	.24
	NPV	.42	.30	.40	.36	.29
	PPV	.91	.86	.79	.78	.76
	ρ	.82	.83	.74	.75	.75
	τ	.74	.78	.75	.79	.80

Note: Phoneme segmentation fluency (PSF) cut scores were based on SAT–10 or OAKS criterion values at the 60th percentile for target, 40th percentile for some risk, and 20th percentile for high risk. *A* represents the area under the ROC curve; NPV is the negative predictive value; PPV is the positive predictive value; ρ is the base rate; and τ is the proportion screened positive (scored below the decision threshold). Thresholds bolded if $A \geq .75$. The 95% confidence intervals were less than ± .02 for all *A* values and ± .01 for all sensitivity and specificity values, except for specificity for the at-risk decision threshold in the fall of Grade 1 (± .02).

Table 4

Optimal NWF Cut Scores for English-Language-Proficient Students

		Kindergarten		1st			2nd
	Statistic	W	S	F	W	S	F
At Risk	*A*	.85	.84	.84	.87	.84	.82
	Threshold	**14**	**34**	**19**	**48**	**62**	**52**
	Sensitivity	.82	.81	.80	.81	.80	.81
	Specificity	.72	.67	.71	.69	.71	.65
	NPV	.83	.81	.87	.87	.86	.87
	PPV	.69	.67	.60	.60	.61	.54
	ρ	.44	.45	.35	.36	.37	.34
	τ	.52	.54	.47	.49	.48	.51
Some Risk	*A*	.87	.86	.84	.83	.82	.79
	Threshold	**19**	**39**	**25**	**54**	**71**	**62**
	Sensitivity	.81	.80	.81	.80	.80	.81
	Specificity	.76	.73	.69	.68	.69	.61
	NPV	.63	.61	.73	.71	.70	.70
	PPV	.89	.88	.78	.78	.79	.74
	ρ	.70	.71	.58	.59	.59	.58
	τ	.64	.65	.60	.61	.60	.63
Target	*A*	.89	.88	.83	.82	.82	.79
	Threshold	**22**	**42**	**30**	**59**	**81**	**70**
	Sensitivity	.80	.80	.81	.81	.80	.80
	Specificity	.81	.76	.67	.66	.67	.60
	NPV	.47	.45	.55	.53	.53	.50
	PPV	.95	.94	.88	.88	.88	.86
	ρ	.82	.83	.74	.75	.75	.75
	τ	.70	.71	.68	.69	.68	.70

Note: Nonsense word fluency (NWF) cut scores were based on SAT–10 or OAKS criterion values at the 60th percentile for target, 40th percentile for some risk, and 20th percentile for high risk. *A* represents the area under the ROC curve; NPV is the negative predictive value; PPV is the positive predictive value; ρ is the base rate; and τ is the proportion screened positive (scored below the decision threshold). Thresholds bolded if $A \geq .75$. The 95% confidence intervals were ± .01 for all *A* values, except for target in the fall of Grade 2 (± .02), and all sensitivity and specificity values, except for specificity for the all decision thresholds in the fall of Grade 2 (± .02).

Table 5

Optimal ORF Cut Scores for English-Language-Proficient Students

		1st		2nd			3rd		
	Statistic	W	S	F	W	S	F	W	S
At Risk	*A*	.92	.95	.89	.91	.91	.84	.85	.84
	Threshold	**13**	**31**	**28**	**55**	**75**	**57**	**76**	**97**
	Sensitivity	.82	.81	.80	.81	.80	.80	.81	.80
	Specificity	.86	.91	.82	.85	.85	.71	.71	.71
	NPV	.90	.90	.89	.89	.89	.92	.92	.92
	PPV	.77	.84	.69	.74	.75	.46	.48	.48
	ρ	.36	.37	.34	.35	.35	.24	.25	.25
	τ	.39	.35	.40	.38	.38	.42	.42	.41
Some Risk	*A*	.91	.93	.86	.88	.87	.80	.82	.81
	Threshold	**19**	**47**	**41**	**76**	**96**	**72**	**89**	**110**
	Sensitivity	.81	.80	.80	.81	.80	.81	.80	.80
	Specificity	.86	.90	.74	.76	.75	.63	.67	.66
	NPV	.76	.76	.73	.74	.72	.80	.80	.80
	PPV	.89	.92	.81	.83	.83	.64	.67	.66
	ρ	.59	.59	.58	.59	.60	.45	.46	.46
	τ	.54	.52	.58	.58	.58	.57	.55	.56
Target	*A*	.90	.91	.85	.87	.86	.80	.81	.81
	Threshold	**26**	**59**	**50**	**86**	**105**	**80**	**100**	**118**
	Sensitivity	.81	.80	.80	.80	.81	.81	.81	.80
	Specificity	.84	.86	.70	.75	.73	.63	.65	.65
	NPV	.59	.59	.54	.55	.54	.62	.63	.62
	PPV	.94	.95	.89	.91	.91	.81	.82	.82
	ρ	.75	.75	.75	.76	.76	.66	.67	.67
	τ	.65	.64	.68	.67	.68	.66	.66	.65

Note: Oral reading fluency (ORF) cut scores were based on SAT–10 or OAKS criterion values at the 60th percentile for target, 40th percentile for some risk, and 20th percentile for high risk. *A* represents the area under the ROC curve; NPV is the negative predictive value; PPV is the positive predictive value; ρ is the base rate; and τ is the proportion screened positive (scored below the decision threshold). Thresholds bolded if $A \geq .75$. The 95% confidence intervals were less than ± .02 for all *A* values and ± .01 for all sensitivity and specificity values.