Hammill Institute on Disabilities

Learning Disability Quarterly
2021, Vol. 44(2) 96–109

DOI: 10.1177/0731948719864408
journals.sagepub.com/home/ldq
SAGE

# Comparison of Literacy Screener Risk Selection Between English Proficient Students and English Learners

**Kelli D. Cummings, PhD[1], Keith Smolkowski, PhD[2], and Doris Luft Baker, PhD[3]**

**Abstract**

Universal screening is a critical component of school-based prevention systems. Screening data enable educators to target students for supplemental intervention, align resources to meet needs, and identify students who may be at risk for learning disabilities. One major requirement of the screening process is that all students are included to gain an accurate picture of school performance. It is therefore surprising that few evaluations of screening systems have focused on English language measures and their use with English learners. In this article, we aim to evaluate common screening thresholds, 54 across Grades k–3, to determine the extent to which they may differ between English learners and English proficient students. Results indicate that many thresholds are consistent between groups with some exceptions in kindergarten. We discuss implications for screening assessment and decision making but suggest that similar cut scores across groups do not imply similar intervention strategies.

**Keywords**

CBM, screening and classification, reading, English language learners

Academic screeners offer educators an important first step in the process of identifying students with learning disabilities. Few literacy screeners, however, have been developed with English learners (ELs) in mind. Most often, developers have designed their screeners, evaluated them, and established cut scores (decision thresholds) with samples that consist of primarily, if not entirely, English proficient students (EPs). Questions have therefore arisen about the use of these screeners for ELs (Sandberg & Reschly, 2011). How do teachers screen for risk of reading disability among ELs with measures developed and decision thresholds defined for EPs? How should teachers interpret the results from those screeners? Students' language proficiency in English and their native language as well as their degree of acculturation appear to adversely affect the validity of some cognitive or standardized achievement tests (Klingner, Artiles, & Méndez Barletta, 2006; Sandberg & Reschly, 2011), which may lead to interpretation problems and, ultimately, misrepresentation in special education (Rueda & Windmueller, 2006). Although the same effects may also be true for screening tools, comparatively few studies have examined the technical adequacy of screeners for EL populations. Those that have vary in their findings, which we describe next.

This article contributes to the burgeoning literature on screening systems for ELs. We first critically examined the limited literature base available, from which only a subset of manuscripts provided valid and generalizable results. Then, we present a study that included 3,418 ELs in Grades K to 3 with 1,174 to 1,970 students per grade and compared them with Smolkowski and Cummings (2016), who conducted a similar analysis with only EPs to reduce the potential influence of language proficiency. Both Smolkowski and Cummings (2016) and the present study used identical signal detection methods and decision rules, and both evaluated the *Dynamic Indicators of Basic Early Skills* (DIBELS), 6th edition (D6; Good & Kaminski, 2002), valid and reliable tools designed for screening reading performance in the early grade levels (Smolkowski & Cummings, 2016). These similarities allow for direct comparisons of results between studies.

## Universal Screening for Risk With ELs

The National Center on Response to Intervention (NCRTI; 2010) recommends, as an essential component of tiered

[1]University of Maryland, College Park, USA
[2]Oregon Research Institute, Eugene, USA
[3]Southern Methodist University, Dallas, USA

**Corresponding Author:**
Kelli D. Cummings, University of Maryland, 3214 Benjamin Building, College Park, MD 20742-1125, USA.
Email: kellic@umd.edu

systems, that schools universally screen students to help identify students with learning disabilities. NCRTI suggests screening with brief assessments or curriculum-based measures (CBMs) with evidence of reliability, validity, and classification accuracy. Reading screeners are widely available, but many have been either developed or normed without including ELs in the sample or without examining the screener performance differences between ELs and EPs (Sandberg & Reschly, 2011). Despite limited evidence, researchers have recommended that educators use English measures with ELs as long as they are learning to read in English (Good et al., 2011). Given the growing EL population (U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics 2016), most of whom are learning to read in English and expected to meet expectations on comprehensive reading tests in English, it is important to understand the accuracy of English CBMs for screening this subgroup.

The few studies that have examined CBM reading measures in populations of ELs often differ in their conclusions. Several studies found comparable psychometric properties and similar predictive accuracy for ELs and EPs with foundational reading screeners, such as measures of phonemic awareness, knowledge of the alphabetic principle, and oral reading fluency (ORF; for example, Betts et al., 2008; Roehrig, Petscher, Nettles, Hudson, & Torgesen, 2008; Sandberg & Reschly, 2011). Vanderwood, Tung, and Checca (2014), however, found that predictive accuracy differed between EPs and ELs, and both Vanderwood et al. and Kim, Vanderwood, and Lee (2016) found that predictive accuracy varied by language proficiency among ELs. Several other papers have also offered evidence for different decision thresholds for ELs and EPs (Hosp, Hosp, & Dole, 2011; Johnson, Jenkins, Petscher, & Catts, 2009; Scheffel, Lefly, & Houser, 2012). In the next section, we summarize the major findings from these papers and highlight key variations between studies. These differences between studies stem from an array of factors, such as differences in the measurement of screener accuracy, reporting of sensitivity and specificity, the measures in use, and the study designs. We also discuss, for a few studies, methodological flaws, including restriction of range, issues of bias, and lack of precision.

## Accuracy

Most studies evaluate the diagnostic accuracy of screeners with methods associated with signal detection theory. These methods typically involve the estimation of a receiver operating characteristic (ROC) curve. The ROC curve is a plot of the true-positive fraction (TPF or sensitivity) against the false-positive fraction (FPF or 1 – specificity) each over the unit interval [0, 1]. Sensitivity "reflects how precisely a screener can detect students with true reading difficulties" (Smolkowski & Cummings, 2015, p. 44). Out of all students who did not meet expectations on the criterion test, sensitivity describes the proportion accurately identified as such by a given screener score (the true positives). Specificity, in contrast, indicates the proportion of all students who met expectations on the criterion test who also screened negative (the true negatives).

The most common summary statistic is the area under the ROC curve (AUC), which captures the overall accuracy of a screener (Cummings & Smolkowski, 2015; Smolkowski & Cummings, 2015; Smolkowski, Cummings, & Stryker, 2016). The AUC captures the trade-off between sensitivity and specificity, and can range from 0.5 for an uninformative screener to 1.0 for a perfect screener. High AUC values imply that for at least some scores on the screener, both sensitivity and specificity also have high values. The AUC describes the accuracy of an entire screener, not unlike the correlation coefficient (technically, the point-biserial correlation; Rice & Harris, 2005). While the ROC curve depicts the overall accuracy as well as trade-offs between sensitivity and specificity values across the range of screener scores, the AUC summarizes the curves and sensitivity-specificity trade-offs with a single value.

## Reporting Sensitivity and Specificity

Typically, at the lowest screener score, sensitivity equals 0 and specificity equals 1, and at the highest screener score, sensitivity equals 1 and specificity equals 0. As sensitivity increases, specificity decreases (see Note 1) and vice versa; they trade off over the range of screener scores. Each pair of sensitivity–specificity values also corresponds to a specific cut score. Sensitivity, specificity, and cut scores all move together, so comments about sensitivity or specificity should always mention all three values, which we call herein the *triplet-reporting requirement* for diagnostic statistics. For example, a screener, such as D6 ORF at the spring of Grade 1, may produce a sensitivity value of .81 and a specificity value of .90 at a cut score of 48 when compared with a particular criterion, such as the 40th percentile on the *Stanford Achievement Test–10th edition* (SAT10; Harcourt Educational Measurement, 2002). A different cut score produces different values of sensitivity and specificity. General comparisons among the statistics, then, such as "sensitivity is higher across the board for English language learner (ELL) students, while specificity is lower" (Scheffel et al., 2012, p. 90; see Note 2), only demonstrate the inherent trade-offs between sensitivity and specificity. The generalization offers no help with interpretation because a different selection of cut points could have produced lower sensitivity and higher specificity. Scheffel and colleagues' subsequent conclusion that "the tests are better at predicting 'at risk students' when they are ELL and are better at predicting 'low risk students' when they are not" (p. 90) apply only to the specific cut scores and not "the tests" themselves. The relative level of sensitivity and

specificity is arbitrary when discussed without reference to the specific cut points. Higher sensitivity and lower specificity for one group when compared with another says nothing about the overall accuracy of the screener for either group because different cut scores for one of the groups (e.g., ELLs) would change those conclusions that are supposedly about the tests overall. Wholesale evaluations of screener performance—evaluations of the screener itself—should therefore depend on the ROC curves and the AUC values because they summarize the whole screener and do not depend on a selected cut score.

## Measure Differences

Several authors have compared their results with those of other studies but without recognizing the differences in screeners and criterion tests. For example, Kim et al. (2016) reported the following:

> Results from the current study are contrary to findings for a third-grade EL group in a study by Hosp et al. (2011). Hosp et al. found much higher sensitivity rates when examining R-CBM in relation to a state test for EL students in third grade. Similarly, Shapiro et al. (2008) also found higher sensitivity levels than specificity levels on [DIBELS ORF] for third-grade students. (p. 15)

These statements violate the triplet-reporting requirement, and refer to different editions of DIBELS and criterion tests. Neither Hosp et al. (2011) nor Shapiro, Solari, and Petscher (2008) explicitly stated which version of DIBELS they used, but they appear to have examined the 6th and 5th editions, respectively, as well as different criterion tests. Kim et al. reported on DIBELS Next (now called Acadience) with yet another criterion test, which makes the comparison questionable. Some screening systems, such as DIBELS 5th, 6th, and Next, also have multiple sets of cut scores to indicate levels of risk (University of Oregon, Center on Teaching and Learning, 2012), and the criterion performance levels on comprehensive tests often represent different percentiles for the population. Comparisons between studies can therefore be difficult, if not impossible, if source manuscripts offer limited details about screeners, decision thresholds, and criterion tests used in their studies. Conceptual generalization may be possible for EL–EP differences but only across studies that provide those details to allow readers to assess how differences in the screeners and criterion measures might affect results, and if studies use similar research designs and statistical methods.

## Study Designs

In most examinations of reading screeners, authors compare subgroups, such as ELs versus EPs, on reliability, growth rates, or predictive validity. Few address decision thresholds through a signal detection approach. Of those, most focus on sensitivity and specificity, with fewer providing ROC curves or AUC values. Other studies on EL–EP differences have not actually addressed the overall diagnostic accuracy of the measures at all, even though they appear to do so. Kim et al. (2016), Scheffel et al. (2012), and Vanderwood et al. (2014) compared ELs at different levels of language proficiency. These authors have actually evaluated a two-stage screening system, where they first selected students by their level of language ability and then examined the diagnostic accuracy within subgroups. This is not how most diagnostic systems are used, nor do the authors compare their results with similar two-stage systems. The approach also potentially leads to *spectrum bias* (Pepe, 2003), when students do not represent the population on the criterion test, as well as other challenges discussed next.

## Range Restriction

Selecting subsets of a sample based on a measure highly associated with reading in English, such as English language proficiency, can restrict the range of possible scores. Range restriction can substantially reduce correlations (Cohen, Cohen, West, & Aiken, 2003), such as those used to demonstrate predictive validity. The principle becomes clear for the correlation coefficient when defined as a function of the covariance and standard deviations: $r_{X,Y} = \mathrm{Cov}_{X,Y}/\sigma_X\sigma_Y$, where $\mathrm{Cov}_{X,Y} = \mathrm{E}[(X-\mu_X)(Y-\mu_Y)]$. Correlation differences may stem from a difference in the covariance or the standard deviations. Kim et al. (2016) and Vanderwood et al. (2014) break their samples into three English proficiency subgroups, which for brevity we call beginning, intermediate, and advanced. In doing so, the standard deviations for ORF and the criterion measures decreased, which reduces the denominator. Reduced standard deviations also imply a smaller covariance, $\mathrm{E}[(X-\mu_X)(Y-\mu_Y)]$, due to the reduced range of scores within each subgroup. Because the standard deviation, $\sqrt{\Sigma(X-\mu_X)/(N-1)}$, is scaled by the sample size, the numerator decreases more dramatically than the denominator in the correlation formula.

Range restriction introduced by subgroup selection can also adversely affect the results based on signal detection theory. The selection of subgroups changes the distributions, and in particular, the base rates differed between subgroups rather dramatically, from 0.51 to 1.0 in Kim et al. (2016). The change in base rates makes predictive values incomparable (Cummings & Smolkowski, 2015; Smolkowski & Cummings, 2015). When Kim et al. sliced their EL sample into subgroups, they also produced a beginner group with no students who met expectations on the state test. A minimum condition for these analyses is that the sample includes enough students who fall into both categories determined by the criterion (e.g., *cases* who score

below expectations and *controls* who meet expectations; Pepe, 2003) that investigators can make generalizations. A study without both cases and controls, or with very few, is not valid. For example, Kim et al. reported specificity of 0 for the beginner sample, but with no controls, specificity equals 0 ÷ 0 and is indeterminate. Even in the intermediate and advanced groups in these studies, the sensitivity and specificity become artifacts of the restricted range on the screeners and criterion tests. Recall that sensitivity represents the proportion of students who scored below the decision threshold on the screener (true positives sample, $N_{TP}$) out of all students who scored below expectations on the criterion test (reading-difficulty sample, $N_D$): $N_{TP}/N_D$. Any procedure that differentially reduces $N_{TP}$ or $N_D$ will change sensitivity (and similarly for specificity). Kim et al. (2016) began with $N_D = 439$ and $N_{TP} = 239$, overall, so sensitivity of 239/439 = .54. The selection of the beginning group reduced $N_{TP}$ by 65% (155) but $N_D$ by 72% (317), so the creation of subgroups based on a measure correlated with the screener and the criterion measure necessarily changed the diagnostic statistics, a manner that makes the incomparable to each other and other studies.

Scheffel et al. (2012) restricted the range in a different way. They examined at-risk and some-risk decision thresholds DIBELS ORF, which is common for DIBELS. But when they computed sensitivity and specificity, they removed the "students classified as 'at some risk'" (p. 85), a procedure attributed to Good, Gruba, and Kaminski (2001). For estimates of sensitivity and specificity, the procedure reduced the denominator sample size (e.g., $N_D$ for specificity) but not the numerator, which gives incorrect estimates. This implies that the Scheffel et al. may have also calculated ROC curves and the AUC incorrectly.

## Bias

The American Educational Research Association (AERA), American Psychological Association, and National Council on Measurement in Education (1999) define predictive bias as "the systematic under- or over-prediction of criterion performance for people belonging to groups differentiated by characteristics not relevant to criterion performance" (p. 179). The papers that discussed bias attended to differences in prediction but not the other two criteria. The under- or over-prediction must be systematic, which requires replication, and the grouping variable that produced differential prediction must have been irrelevant to criterion performance, which is certainly not the case for comparing groups that differ on English language proficiency on reading assessments. Differential prediction, also called differential response or moderation, is often not a bias at all. Although differential prediction may be due to some form of bias, it may stem from one or more of many other relevant factors that can affect prediction (e.g., quality of instruction, range restriction, vocabulary). Due to the differences in Spanish and English orthographies, Spanish-speaking ELs may learn to decode at a different rate than EPs, which could in turn change the associations between a measure like D6 nonsense word fluency (NWF) and a comprehensive reading assessment (e.g., Fien et al., 2008). But because the characteristics that differ between ELs and EPs are relevant to reading performance, they do not imply bias. Nonetheless, several papers have incorrectly labeled or suggested that differences in correlations between two groups represent bias (e.g., Betts et al., 2008; Hosp et al., 2011; Kim et al., 2016; Sandberg & Reschly, 2011; Vanderwood et al., 2014) without considering all criteria.

## Precision

Few authors estimated precision, and most authors reported sensitivity and specificity values as if they were precise estimates. Of the papers reviewed, only Kim et al. (2016) provided confidence bounds, which was laudable, and demonstrated the need for confidence intervals. Due to its relatively small samples ($N < 300$), many of the bounds spanned 20% to 40% of the range in sensitivity or specificity. Kim et al. discussed these values as if they were precise, such as "this study found 69% sensitivity" (p. 15). In contrast, the confidence interval [.47, .61] tells us that, if the study were repeated many times and all model assumptions were correct, 95% of the confidence intervals from the repeated studies would contain the true estimate (Greenland et al., 2016). Nonetheless, Kim et al. at least provided the estimates with their results. No other study offered this important information about the coverage probability of their estimates, and many treated their estimates as if they contained little error. As with all statistics, authors should report confidence intervals for the AUC, sensitivity, and specificity.

## Summary

After accounting for the various challenges above, we focused on the six studies that remained relevant to the comparison between ELs and EPs on literacy screeners. Some of the literature examined reliability and predictive validity, which were less relevant to the present article as those that used signal detection theory methods but nonetheless reflected the differences and similarities in screener performance between ELs and EPs. Our primary emphasis was on studies that addressed diagnostic accuracy.

Betts et al. (2008) compared the *Minneapolis Kindergarten Assessment* (MKA), which includes several fluency screeners for literacy, with the Northwest Evaluation Association (2004) standardized test of reading. With a reasonably large sample, 544 ELs and 1,375 EPs, this study did not find differences in predictive validity between groups

after accounting for economic disadvantage defined by free or reduced-price lunch status (cf. Meehl, 1971). De Ramirez and Shapiro (2006) investigated the growth rates between ELs and EPs on AIMSweb ORF on a very small sample. They found that CBMs could be valuable for teachers who make instructional decisions for ELs, although expected growth rates may need to differ. De Ramirez and Shapiro did not assess validity, and none of these authors addressed screener decision thresholds. Roehrig et al. (2008) examined DIBELS ORF, 5th edition, when predicting the Florida state test and found no evidence of differential predictive validity. Although they estimated ROC curves and related statistics, they did not use them to compare ELs and EPs.

Hosp et al. (2011) compared several pairs of groups (e.g., economic disadvantage vs. not, disability status vs. not), including limited English proficiency versus not, on the some-risk benchmark cut scores for D6 NWF and ORF in Grades 1, 2, and 3. They found statistically significant differences ($p < .001$) between ELs and EPs on sensitivity for two assessments, ORF in the winter and spring of Grade 3, specificity for seven measures, NWF and ORF in the winter of Grade 1, and ORF from the winter of Grade 2 through the spring of Grade 3. They found no differences with the AUC (see Note 3), which were all above .80 for ORF and all below .80 for NWF. The absence of AUC differences implies that the sensitivity and specificity differences involved trade-offs.

Johnson et al. (2009) explored the selection of new cut scores for ELs and EPs, as well as students who received free and reduced-price lunch, for NWF in the spring of kindergarten and ORF in the fall of Grade 1 from DIBELS 5th edition. They selected cut scores when sensitivity equaled .90, which produced decision threshold 4 to 5 points lower for ELs than for comparison students. The AUC values were also lower for ELs than for comparison students. The comparison group, however, included students who were neither ELs nor qualified for free or reduced-price lunch, which made it difficult to generalize from the results in Johnson et al.

The present study replicates the methods of Smolkowski and Cummings (2016), who selected a different set of decision thresholds for EPs than those recommended by Good and Kaminski (2002); Good, Simmons, and Kame'enui (2001); and related sources. Significant ambiguity surrounded the specification of initial cut scores for D6. Good and Kaminski reported odds of assigning students to categories, but the odds were not defined and appeared to represent proportions, percentages (e.g., "the odds are 56 percent," p. 57), or predictive values. Many of the decision thresholds were also based on comparisons with other D6 measures administered later in the same year rather than a criterion test, which introduces bias into the system (Smolkowski & Cummings, 2015). Due to the uncertainty surrounding the methods used to set decision thresholds, Smolkowski and Cummings (2016) first estimated ROC curves and the AUC and then established cut scores when sensitivity equaled .80, with 4,600 to 5,600 EPs per grade level. They reported the AUC, decision thresholds, sensitivity, and specificity, among other statistics, for each measure from kindergarten to Grade 3. The present study parallels the Smolkowski and Cummings (2016) but with a sample of ELs rather than EPs to offer guidance to educators interested in selecting students at risk for reading disability.

## Research Aims

This study addresses two primary research questions:

> **Research Question 1:** What are the optimal screening decision thresholds for ELs?
> **Research Question 2:** Do they meaningfully differ between ELs and EPs?

To test these questions, we examined the extent to which D6 measures accurately predicted criterion scores at the 20th and 40th percentiles on comprehensive reading measures in the spring of the same year in which D6 measures were administered. The lower level implies risk of reading disability. We then selected optimal cut scores for ELs and examined their classification performance. Beyond the traditional thresholds for students who are at risk or have some risk, we identified a *target* level of performance, associated with the 60th percentile on comprehensive reading measures (Smolkowski et al., 2016). Because educators naturally anchor expectations (Jacowitz & Kahneman, 1995), we suggested target cut scores so educators could set higher goals for their students who have already demonstrated minimal proficiency (i.e., benchmark). Finally, we compared optimal thresholds for ELs with those set for EPs in Smolkowski and Cummings (2016).

Although we used rigorous methods for this evaluation, in certain respects, we chose to emphasize generalizability to authentic settings and the practical value for educators. For example, we chose to define EL as a student who had received services for limited English proficiency in the participating schools. Although this definition has less validity than a carefully administered, standardized language measure, it is more practical and generalizable because the determination matches the assignment practices in many schools. This sample also allowed us to focus on students who, at the time of assessment, required language supports. We therefore excluded students for whom English may not have been their first language but who no longer required support by their schools. We also chose to compare cut scores between ELs and EPs rather than diagnostic statistics with the goal that this information would be more readily accessible to practitioners without requiring a background in this type of methodology. We discuss the underlying

**Table 1.** Descriptive Information for English Learners.

| Grade | Measure | Fall | | | Winter | | | Spring | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | *n* | *M* | *SD* | *n* | *M* | *SD* | *n* | *M* | *SD* |
| K | LNF | 1,180 | 3.4 | 7.3 | 1,292 | 16.5 | 14.7 | 1,361 | 29.3 | 17.3 |
| | PSF | | | | 1,292 | 14.4 | 14.1 | 1,361 | 36.0 | 17.6 |
| | NWF | | | | 1,288 | 8.0 | 10.0 | 1,361 | 23.9 | 17.3 |
| | SAT10 percentile | | | | | | | 1,370 | 17.6 | 17.8 |
| Grade 1 | LNF | 1,828 | 24.4 | 16.5 | | | | | | |
| | PSF | 1,828 | 24.6 | 17.7 | 1,907 | 41.9 | 16.6 | 1,960 | 48.2 | 13.4 |
| | NWF | 1,827 | 17.0 | 17.7 | 1,907 | 44.7 | 24.1 | 1,960 | 63.0 | 29.4 |
| | ORF | | | | 1,907 | 16.8 | 20.1 | 1,960 | 37.5 | 27.9 |
| | SAT10 percentile | | | | | | | 1,970 | 19.9 | 19.9 |
| Grade 2 | NWF | 1,655 | 50.0 | 30.5 | | | | | | |
| | ORF | 1,670 | 30.4 | 25.0 | 1,734 | 54.7 | 34.4 | 1,774 | 73.7 | 37.5 |
| | SAT10 percentile | | | | | | | 1,796 | 20.3 | 19.7 |
| Grade 3 | ORF | 1,174 | 55.0 | 29.8 | 1,202 | 71.9 | 34.6 | 1,220 | 91.9 | 35.5 |
| | OAKS raw score | | | | | | | 1,244 | 205.3 | 8.8 |

*Note*. For the OAKS, raw scores are reported; percentiles were unavailable. LNF = letter naming fluency; PSF = phoneme segmentation fluency; NWF = nonsense word fluency; SAT10 = Stanford Achievement Test, 10th edition; ORF = oral reading fluency; OAKS = Oregon Assessment of Knowledge and Skills.

statistics and provide details in the Supplemental Material, but group differences in the recommended cut scores would have the most practical value.

## Method

Data for this study were collected over a period of 3 years, from the fall of 2003 to the spring of 2006 from first grade to third grade, in a sample of students from 34 Oregon Reading First (ORRF) schools in 16 independent school districts (see Baker et al., 2011). School teams collected D6 data from 2003–2004 to 2005–2006 and administered the SAT10 at the end of each year for students in kindergarten Grades 1 and 2. Those same teams administered the *Oregon Assessment of Knowledge and Skills* (OAKS; Oregon Department of Education [ODE], 2008) section on reading and literature for students at the end of Grade 3.

### *Participants and Setting*

Participants were 3,418 ELs attending 34 schools in the Pacific Northwest. Half of the schools were located in urban areas and the rest were roughly equally divided between mid-size cities and rural areas. For the present study, we defined ELs as students who received services for limited English proficiency from at some point between kindergarten and Grade 3. The schools utilized state criteria to designate ELs, which included two primary components, a home language survey that indicated English was not the primary language spoken in the home and demonstration of a lack of or limited English proficiency defined by a standardized measure (Fien et al., 2008). About 20% of the total sample met this definition and provided scores on Oregon's end-of-year comprehensive reading tests. Students who did not require services for limited English proficiency did not participate in the present sample, but Smolkowski and Cummings (2016) included them in their analysis of EPs. As students provided data across multiple years, the final sample included 6,380 criterion test scores (see Table 1).

Half of the students, 48%, were female and students fell into the following racial-ethnic categories: 14% Caucasian, 68% Hispanic or Latinx, 2% African American, 6% American Indian, 9% Asian, and less than 1% Alaskan Native, Hawaiian, Pacific Islander, or Other. Statewide, more than two thirds of ELs were Hispanic or Latino and spoke Spanish, the next most common languages were Russian (4%) and Vietnamese (2.6%). About 10% of students in the involved districts received special education services. Among ORRF schools, 69% of students qualified for free or reduced-price lunch, and 27% of third graders did not meet minimum expectations on the OAKS during the course of the study. For additional details, see Baker et al. (2008, Baker et al., 2011).

Students were administered D6 measures in the fall, winter, and spring of Grades K through 3 and the two criterion measures, the SAT10 or OAKS, in the spring of each year. Of those students who completed comprehensive reading tests, 86% to 99% of students in Grades K through 3 participated in the D6 assessments. Students were administered comprehensive reading tests in the spring, and on average, less than 4% were excluded due to absences.

## Criterion Measures

*SAT10.* The SAT10 (Harcourt Educational Measurement, 2002) is a group-administered, norm-referenced test of overall reading proficiency. The measure is untimed. 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 total reading score as our criterion with 2007 norms based on a representative sample of the U.S. student population.

*OAKS.* The OAKS, developed by the ODE (2008), is an untimed, multiple-choice test administered yearly to all Grade 3 students in Oregon. Reading passages represented literary, informative, and practical selections that students might encounter in school settings and other reading activities. Individual subtests require students to understand word meanings in the context of a selection; locate information in common resources; answer literal, inferential, and evaluative comprehension questions; recognize common literary forms, such as novels, short stories, poetry, and folk tales; and analyze the use of literary elements and devices, such as plot, setting, personification, and metaphor. ODE reported criterion validity of .75 with the *California Achievement Tests* and .78 with the *Iowa Tests of Basic Skills*. The scores from the four alternate test forms used for the OAKS demonstrated Kuder–Richardson reliability of .95.

## D6

Dynamic Measurement Group (DMG, 2008) summarized test–retest and alternate-form reliability and concurrent and predictive validity estimates for D6 measures from 26 studies with 29 criterion tests.

*Letter naming fluency (LNF).* LNF measures the number of randomly ordered upper- and lowercase letters students name in 1 min. Score reliabilities ranged from .86 to .98 and validity estimates from .31 to .74 (DMG, 2008).

*Phoneme segmentation fluency (PSF).* PSF measures phonemic awareness; students are scored on the number of correct individual phonemes segmented from words read aloud by the examiner in 1 min. Score reliabilities ranged from .74 to .90 and validity coefficients from .43 to .59 (DMG, 2008).

*NWF.* NWF measures alphabetic understanding and phonological recoding ability (Cummings, Dewey, Latimer, & Good, 2011). Students are scored on the number of phonemes they correctly identify from consonant-vowel and consonant-vowel-consonant pseudowords (either individual sounds or whole pseudowords) in 1 min. Score reliabilities ranged from .84 to .98 and validity coefficients from .33 to .82 (DMG, 2008).

*ORF.* DIBELS ORF measures fluency with connected text. Students read sets of three passages, 1 min each, and are scored on the median number of correctly read words. Score reliabilities ranged from .89 to .99 and validity estimates from .31 to .97 (DMG, 2008).

## Data Collection

Data collectors were grouped into teams and received 1-day trainings on D6 administration and scoring with additional calibration sessions from reading coaches at each school. Test–retest reliabilities ranged from .60 to .83 for PSF scores, .83 to .90 for NWF scores, and .93 to .97 for ORF scores. Teachers administered the SAT10 and the OAKS each spring. SAT10 testing was monitored by Reading First coaches trained by the ORRF Center. Coaches trained teaching staff in their building on test administration and monitoring. Coaches documented testing procedures with an 18-item implementation fidelity checklist; median fidelity was 98.3%. Teachers administered the OAKS according to procedures established by the school, district, and state.

## Analysis Approach

These analyses followed the methods outlined in Smolkowski and Cummings (2016). We first generated ROC curves and calculated the AUC, $A$, for each measure administered at each time point to evaluate overall accuracy with respect to end-of-year criterion tests. Values of $A$ above .950 indicate excellent screeners, .850 to .949 suggests very good screeners, and .750 to .849 corresponds to moderately reasonable screeners (Smolkowski & Cummings, 2015; Swets, 1988). Values below .750 represent relatively poor utility, where teacher judgments may be more valuable than the results of a reading screener (Martin & Shapiro, 2011).

The selection of a decision threshold for each level of risk should depend on the anticipated consequences of four potential outcomes: false and true positives and negatives. We set decision thresholds based on the complement of sensitivity, the false-negative fraction, so no more than 20% of students from the reading-difficulty population were incorrectly identified as typically achieving (i.e., sensitivity = .80). Establishing decision thresholds with sensitivity values allows for a consistent interpretation of cut scores for all measures at all assessments. The rationale stems partly from the common practice with DIBELS of multiple decision thresholds to indicate different levels of risk. False negatives for the highest risk level will not lead to the most intensive supports, but such students will likely receive a true-positive score for the *some-risk* threshold and consequently receive supplemental instruction as well as progress monitoring. Such students may be reassigned to receive intensive supports more quickly than those who might otherwise remain in core instruction. Teachers may also be able to identify false positives—typically achieving students incorrectly assigned

**Table 2.** Areas Under the ROC Curve and Decision Threshold for ELs and EPs for DIBELS 6th Edition Measures.

| DIBELS measure | and assessment time | | At risk | | | Some risk | | | Target | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | Threshold | | | Threshold | | | Threshold | |
| | | | *A* | ELs | EPs | *A* | ELs | EPs | *A* | ELs | EPs |
| LNF | K | F | .68 | 3 | **6** | .74 | 4 | **11** | .80 | **4** | **14** |
| | | W | .83 | **21** | **27** | .86 | **27** | **34** | .89 | **28** | **37** |
| | | S | .81 | **38** | **42** | .95 | **42** | **47** | .88 | **43** | **50** |
| | Grade 1 | F | .81 | **29** | **33** | .83 | **34** | **38** | .87 | **36** | **42** |
| PSF | K | W | .75 | 21 | **28** | .76 | **26** | **33** | .79 | **27** | **36** |
| | | S | .73 | 48 | 54 | .75 | 50 | 57 | .75 | 51 | 58 |
| | Grade 1 | F | .71 | 37 | 40 | .71 | 40 | 44 | .74 | 41 | 45 |
| | | W | .67 | 53 | 56 | .65 | 55 | 59 | .65 | 56 | 60 |
| | | S | .60 | 59 | 61 | .58 | 59 | 62 | .59 | 59 | 62 |
| NWF | K | W | .82 | **9** | **14** | .85 | **13** | **19** | .88 | **14** | **22** |
| | | S | .82 | **30** | **34** | .88 | **34** | **39** | .90 | **36** | **42** |
| | Grade 1 | F | .82 | **18** | **19** | .84 | **24** | **25** | .87 | **27** | **30** |
| | | W | .82 | **49** | **48** | .83 | **55** | **54** | .85 | **58** | **59** |
| | | S | .80 | **67** | **62** | .81 | **75** | **71** | .84 | **79** | **81** |
| | Grade 2 | F | .78 | **55** | **52** | .79 | **64** | **62** | .79 | **70** | **70** |
| ORF | Grade 1 | W | .91 | **13** | **13** | .92 | **19** | **19** | .94 | **22** | **26** |
| | | S | .93 | **32** | **31** | .93 | **48** | **47** | .94 | **54** | **59** |
| | Grade 2 | F | .86 | **30** | **28** | .87 | **40** | **41** | .87 | **47** | **50** |
| | | W | .89 | **58** | **55** | .89 | **75** | **76** | .88 | **81** | **86** |
| | | S | .88 | **82** | **75** | .87 | **97** | **96** | .87 | **104** | **105** |
| | Grade 3 | F | .78 | **59** | **57** | .76 | **71** | **72** | .80 | **76** | **80** |
| | | W | .78 | **77** | **76** | .76 | **91** | **89** | .80 | **97** | **100** |
| | | S | .77 | **102** | **97** | .76 | **111** | **110** | .79 | **117** | **118** |

*Note.* *A* represents the area under the ROC curve. Decision thresholds were based on SAT10 or OAKS criterion values for the 20th, 40th, and 60th percentile for at risk, some risk, and target, respectively, and bolded if $A \geq .75$. Smolkowski and Cummings (2016) produced thresholds for EPs, provided here for comparisons with ELs. EL – EP threshold differences ranged from 7 points (ORF, Grade 2, spring, at risk, 82 − 75) to −10 points (LNF, kindergarten, fall, target, 4 – 14) and generally increased from kindergarten to Grade 3. ROC = receiver operating characteristic; ELs = English learners; EPs = English proficient student; DIBELS = Dynamic Indicators of Basic Early Skills; LNF = letter naming fluency; PSF = phoneme segmentation fluency; NWF = nonsense word fluency; ORF = oral reading fluency.

to extra supports—through progress monitoring and reassign those students to core instruction when needed. Our choice of decision thresholds also hinges on the observation that most reading screeners are not highly accurate (i.e., $A < .95$). Setting cut scores based on high sensitivity values, such as .90 or greater, can produce a large number of false positives, which can overwhelm support systems.

We estimate *A* and its standard error with SAS (SAS Institute, 2016) PROC LOGISTIC and other statistics with PROC FREQ. For reporting, we followed the Standards for the Reporting of Diagnostic accuracy studies (STARD; http://www.stard-statement.org/).

## Results

Table 1 provides descriptive information for the SAT10, OAKS, and D6 measures for ELs. Table 2 presents values for the AUC, *A*, and the decision thresholds for each measure as well as the cut scores for EPs produced by Smolkowski and Cummings (2016). We estimated the statistics for students *at risk* with criterion test scores below the 20th normative percentile, at *some risk* (benchmark) for students below the 40th percentile, and *target* for students below the 60th percentile on comprehensive tests. We presented *A*, sensitivity, and specificity with confidence intervals in the Supplemental Material along with estimates of negative predictive value, positive predictive value, the base rate, and the proportion of students who screened positive.

### *Screener Accuracy*

All LNF, NWF, and ORF demonstrated adequate accuracy ($A \geq .75$), with minor exceptions. LNF collected in the fall of kindergarten failed to exhibit adequate accuracy for the at-risk and some-risk criteria. For ORF collected in Grade 1, *A* exceeded .90, demonstrating excellent discrimination between students above and below the end-of-year reading

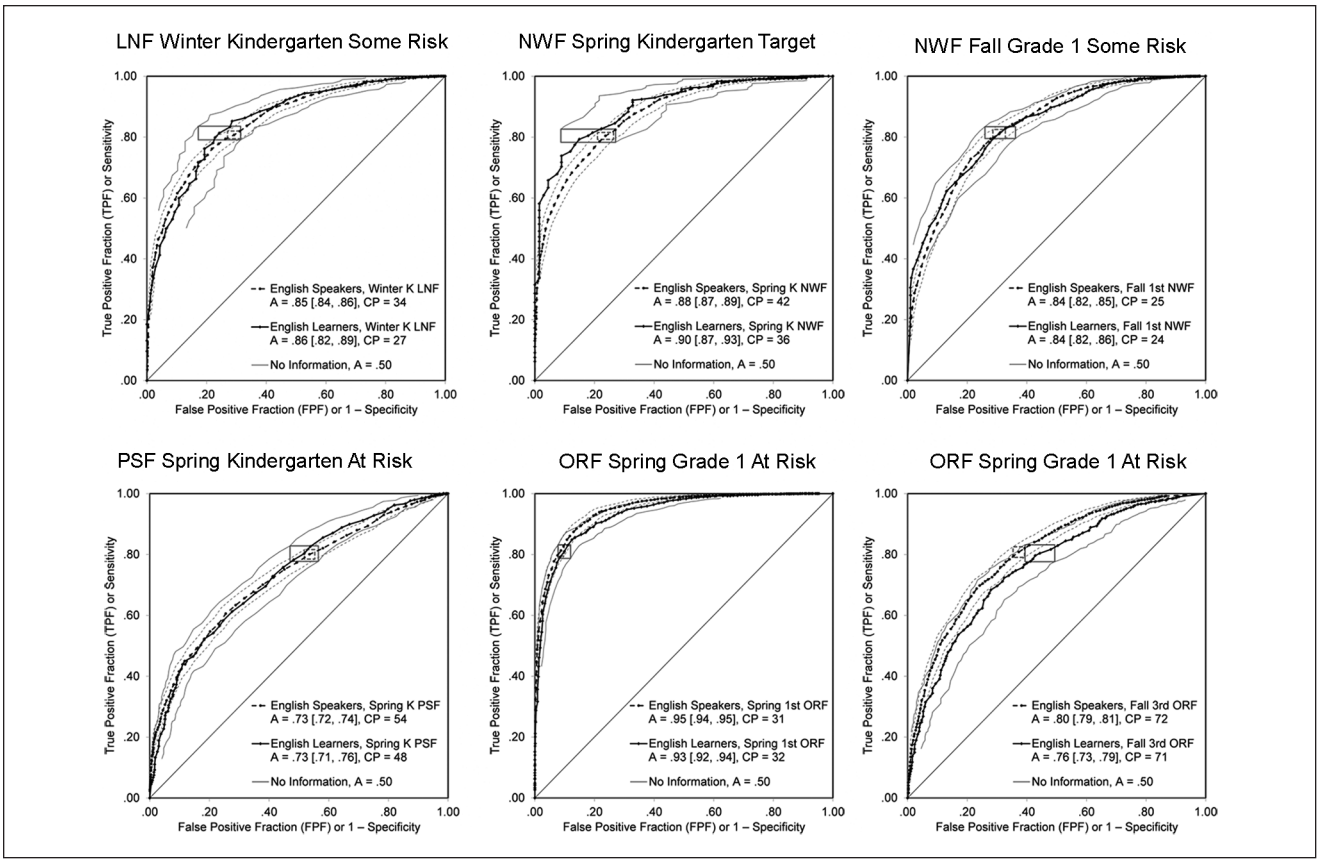


**Figure 1.** ROC curves for six DIBELS assessments.

*Note.* The solid line depicts the ROC curve for ELs; the solid, light outer lines show the 95% confidence bounds around the curve; and the solid, larger box shows the confidence bounds around the chosen cut score (see legend in each graph). The dashed lines show the ROC curve for EPs, with the 95% confidence bounds depicted by light dashed lines and the small dashed box showing the confidence bounds around the chosen cut score. The chosen cut scores were defined as the point with sensitivity closest to .80. Most of differences in cut scores selected for EPs and ELs represent a small change in terms of their location on the ROC curve. In all curves plotted here, except the ORF fall Grade 3, some-risk graph, the confidence bounds around the chosen decision threshold for ELs completely contain the confidence bounds around the decision threshold for EPs. ROC = receiver operating characteristic; DIBELS = Dynamic Indicators of Basic Early Skills; LNF = letter naming fluency; NWF = nonsense word fluency; PSF = phoneme segmentation fluency; ORF = oral reading fluency; ELs = English learners; EPs = English proficient student.

performance criteria. PSF, in contrast, seldom achieved $A \geq .75$, except for the winter of kindergarten, and then only for the some-risk and target decision thresholds. Although $A \geq .75$ for most measures collected at most time points, the confidence intervals in the Supplemental Material indicate that accuracy could fall outside that criterion level in replication samples.

For each administration and risk level, Table 2 presents $A$ and the selected cut score. We selected the lowest score with sensitivity at or above .80 for all measures, so sensitivity values ranged from .80 to .84. Specificity ranged from .46 to .81 for LNF, .24 to .58 for PSF, .60 to .83 for NWF, and .55 to .96 for ORF. Due to the trade-offs between sensitivity and specificity captured by the AUC, $A$, and because we selected scores on sensitivity values, specificity and $A$ should have been highly correlated in our results. We estimated correlations between $A$ and specificity of .99 across all measures and risk levels, with correlations of .99 for ORF, .97 for NWF, .98 for PSF, and .95 for LNF.

## EL–EP Differences

*ROC curves.* Figure 1 shows ROC curves for ELs (solid lines) and EPs (dashed) on a sample of six D6 administrations. We selected the measures, their timing, and risk levels to demonstrate the range of differences, but we only selected one, PSF in the spring of kindergarten, that did not meet our criterion for acceptable overall accuracy ($A < .75$). The figure shows the curves with their confidence bounds across the range of scores with sufficient sample sizes to estimate them. The confidence bounds are narrower for EPs, based on over 4,078 to 5,634 students, than for ELs, estimated with 1,180 to 1,970 students. Each pair of curves includes a box that shows the

confidence bounds for sensitivity and specificity at the selected cut score (see legends for cut scores). For at-risk ORF in the spring of Grade 1, the box for EPs is small and overlaps the box for ELs on two sides (lower left). The rectangular nature of the boxes for ELs—wider than tall—occurred because fewer ELs scored above the criterion level of performance, especially for the target performance of NWF in the spring of kindergarten, which produces wider bounds on specificity.

The ROC curves differ between measures in their overall accuracy. As *A* approaches .90 (e.g., spring Grade 1 ORF), the curves indicate much better accuracy than those with *A* values below .75 (e.g., spring kindergarten PSF). Regardless of the performance differences between measures, the curves for ELs and EPs generally overlap and sometimes cross each other. The confidence bounds for ELs often included the curve for EPs across the range of each screener and especially for those near the upper left corner, where both sensitivity and specificity take on high values. For all curves except ORF in the fall of Grade 3 for some-risk (lower right), the confidence bounds for the EL cut score encompassed the cut score for ELs.

*Decision thresholds.* For the 23 assessment times, we set a decision threshold for each of three risk levels, 69 in total, and then compared them with the cut scores in Smolkowski and Cummings (2016) in Table 2. All but three differed by at least 1 point, 34 differed by 3 points or less, and 52 differed by 5 points or less. Cut score differences ranged from −10 (a lower score for ELs) to +7 (higher score for ELs), but 15 administrations did not meet the minimum *A* value for either ELs or EPs. We set aside those 15 cut scores, with differences from 2 to 7 points, and summarize the differences between ELs and EPs for the remaining 54 next.

More than half the cut scores (28 of 54) differed by 3 points or fewer. The differences in decision thresholds between ELs and EPs were generally greater in kindergarten than later grades. The largest difference of 10 points, for example, occurred for the LNF target level at the fall of kindergarten, and of the 12 differences greater than 5 points, 11 arose for kindergarten measures. For NWF in kindergarten, cut scores were 4 to 8 points lower for ELs than EPs, but for the fall and winter of Grade 1, differences decreased to just ±1 point, except for the fall target threshold at −3 points. In Grades 2 and 3, the differences were smaller and several of the cut scores were higher for ELs than EPs, such as ORF at-risk thresholds in Grades 2 and 3. The differences were generally smaller for the at-risk threshold, larger for the some-risk threshold, and largest for the target threshold. In summary, the largest negative differences were in the top right of Table 2 (–10 for target LNF in the fall of kindergarten) and generally increased to the lower left (+5 for at-risk ORF in the spring of Grade 3), and on average, the cut scores were about 5 to 7 points lower for ELs in kindergarten and roughly the same in Grades 1, 2, and 3.

If teachers used the cut scores produced for EPs with ELs, in most cases, the sensitivity and specificity values would have changed by less than the width of their confidence bounds (±.02 to ±.04). For the at-risk cutoff for ORF at the end of Grade 1, scores differed by just 1 point. If ELs used the EP cut score of 31 rather than 32, sensitivity would decrease from .81 to .79 and specificity would increase from .90 to .91. For a few of the larger differences, such as the at-risk threshold for LNF in the winter of kindergarten, the change in decision threshold from 21 for ELs to 27 for EPs would have changed sensitivity from .81 to .89 and specificity from .65 to .53.

We depicted the decision thresholds for both ELs and EPs in Figure 2 for NWF and ORF across grade levels. The figures demonstrate three notable features. First, the thresholds increased substantially across each school year, although they also dropped slightly during the summer breaks. Second, the decision thresholds for ELs were generally lower than those for EPs for NWF and target levels of ORF, but EL cut points were higher for the at-risk and some-risk cut points for ORF. Third, for NWF and ORF, the thresholds for ELs did not markedly differ from those for EPs. That is, the variability between ELs and EPs was generally much smaller than the variability from one assessment to the next.

## Discussion

We tested the accuracy of D6 for identifying ELs at different levels of risk for reading disability within an effective, research-based, tiered model of reading instruction (Baker et al., 2011). Most D6 measures were accurate overall, except for PSF. The overall accuracy of PSF, indicated by the AUC, was too low for teachers to base decisions on this measure except for the winter of kindergarten. ORF, especially in Grade 1, produced the highest assignment accuracy across risk levels. These results parallel those of Smolkowski and Cummings (2016) for EPs, who concluded that almost all measures except PSF would likely improve teacher decisions.

The ROC curves demonstrated a few differences between ELs and EPs, as the confidence bounds for the curves for ELs mostly contained the curve for EPs, in overall accuracy of the screeners across risk levels. A comparison of AUCs between the present study and Smolkowski and Cummings (2016) also suggests minimal differences. Of the 69 estimates of the AUC for ELs, the confidence bounds for 46 of them (67%) included the AUC for EPs, and of the 23 exceptions, the AUC was slightly higher for ELs 6 times and higher for EPs 17 times. In several of the cases where the AUC for EPs was outside the confidence intervals for ELs, the differences were small (e.g., $A = .91$ for at risk winter

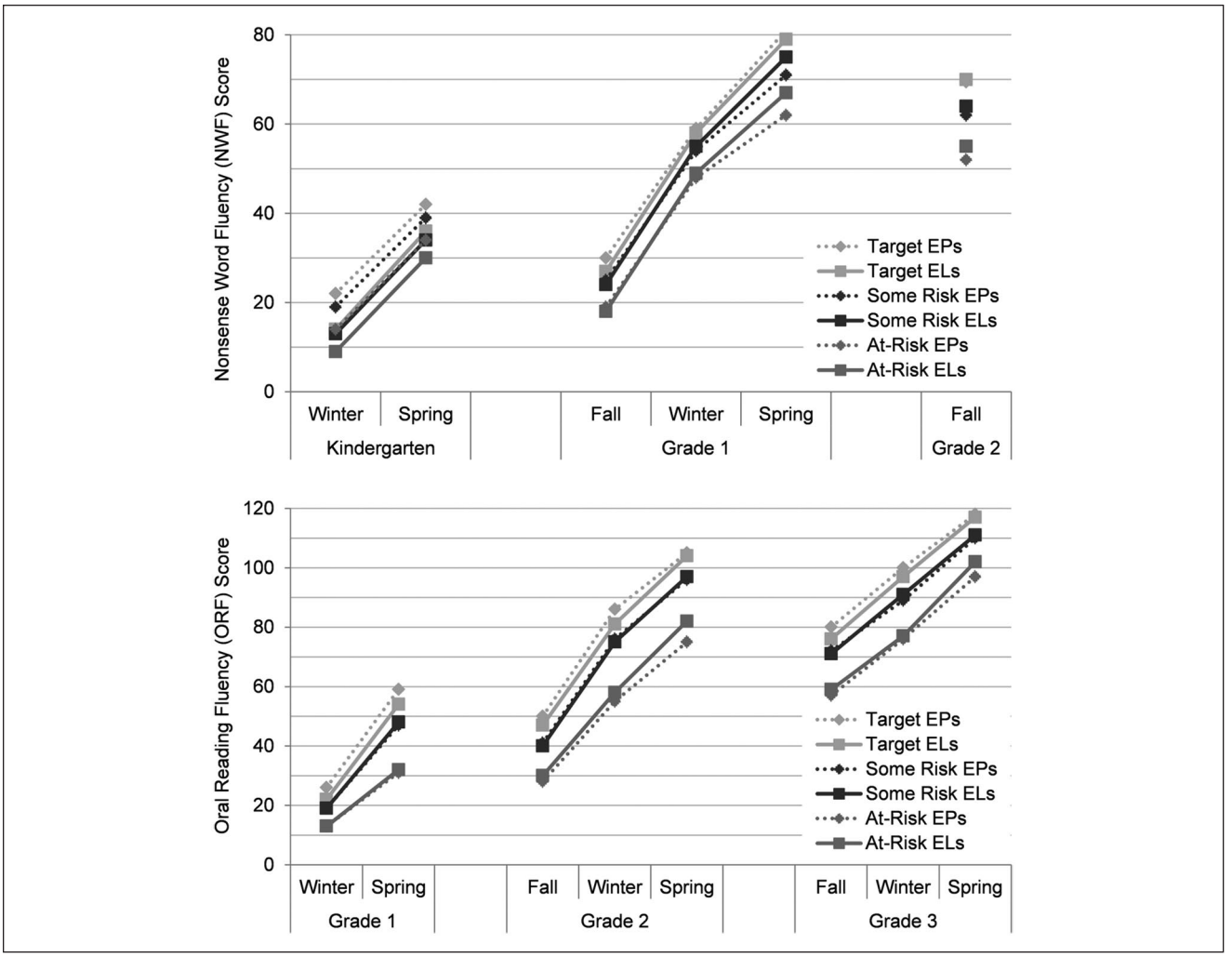


**Figure 2.** NWF (top) and ORF (bottom) decision thresholds for ELs (solid lines) and EPs (dashed lines) for at-risk, some-risk, and target performance levels.

*Note.* The lines for EPs at risk or with some risk in Grade 1 and with some risk in Grade 2 are partially obscured by the respective lines for ELs (near-perfect overlap). Cut scores for EPs are taken from Smolkowski and Cummings (2016). NWF = nonsense word fluency; ORF = oral reading fluency; ELs = English learners; EPs = English proficient student.

Grade 2 ORF for EPs but $A = 89$, [.87, .90], for ELs). Given the interpretation of the AUC—the likelihood of placing two students, one who met expectations on the criterion and one who did not, correctly into risk categories by the screener—these differences have limited practical value. The utility of the measures for use in screening appears to be comparable between ELs and EPs.

The ROC curves demonstrate that the evaluation of diagnostic accuracy with precision requires a large sample (Malhotra & Indrayan, 2010). In Figure 1, the narrow bounds around the ROC curves for EPs relied on over 4,300 students (4,393 to 4,885) for each administration. The confidence bounds for ELs, in contrast, have relatively wider confidence bounds given the smaller sample size, even when estimated with well over 1,100 ELs (1,174 to 1,960). Several studies that examined diagnostic accuracy for ELs included well below 1,000 students (e.g., 247 to 403 in Hosp et al., 2011; 159 to 596 in Vanderwood et al., 2014). Small samples, especially fewer than 250, will produce confidence bounds so wide that reliable comparisons between groups may not be feasible.

Even large samples can also produce relatively jagged ROC curves, such as for the target threshold for spring kindergarten NWF in Figure 1 (top center). The jagged nature of curves may stem from random sample fluctuations or from a single skill (e.g., learning a certain blend) that allows students to answer several more items correctly. For these reasons, a decision based on a single point, especially with a small sample, may miss the bigger picture. Evaluations and subsequent decisions based solely on sensitivity and

specificity for a single point may fall within the general trend of the curve or a point that juts out—an outlying point on the curve. Hence, ROC curves, the AUC, and their confidence bounds are critical features for screener evaluation.

The decision thresholds from the present study did not differ substantively from those established for EPs in a prior study (see Figure 2). Indeed, the difference between decision thresholds for ELs and EPs was often much smaller than the difference between the thresholds for EPs proposed by Smolkowski and Cummings (2016) and the original thresholds for D6 (Good, Simmons, Cummings, et al., 2001, Good, Simmons, Kame'enui, Kaminski, & Wallin, 2002). The differences were greatest in kindergarten, perhaps because of lower English language proficiency or, in part, because of a lack of opportunities for ELs to attend preschool (García, 2015). Alternatively, ELs may have a discrepancy between their *basic knowledge* of English, such as vocabulary and comprehension, and their *fluency* with English. Although students may be able to read and understand a word, they may need additional practice to become fluent (Smolkowski et al., 2016). Such a discrepancy between knowledge and fluency could explain the differences in cut scores achieved, given most standardized tests are not timed, including both criterion tests used in this study. The underlying cause of the discrepancy requires more research.

For educators concerned about the difference, we can also offer a simple rule that would align most cut scores reasonably well: subtract 5 points from the EP decision thresholds for ELs in kindergarten but use the same cut scores for ELs and EPs in Grades 1 through 3. That is, for kindergarten only, ELs may initially score about 5 points lower than their EP peers.

Although the cut scores were similar between ELs and EPs, instruction or interventions should and likely will be different. If schools provide bilingual education, it may be more efficient for ELs to spend time building upon their first language to promote cross-linguistic transfer (Gerber et al., 2004; Nakamoto, Lindsey, & Manis, 2012). Schools that have limited capacity to provide native language instruction could train teachers on the similarities and differences between orthographic systems to understand where ELs may have difficulty in their letter sound recognition (see Honig, Diamond, & Gutlohn, 2000). Of course this would be a challenge for the many ELs who speak languages other than Spanish, which includes nearly one third of the current sample. We argue that cut scores provide information about which children are at risk for reading disability and how severe that risk might be, but they do not prescribe any particular supports or interventions (Smolkowski & Cummings, 2015; Smolkowski et al., 2016). The screeners, such as NWF or ORF, may help teachers identify deficits for ELs as they would for EPs. As argued by Klingner et al. (2006), vocabulary, cultural, and contextual factors likely have a great influence on reading performance for ELs, so educators might also want to examine measures of language and vocabulary to aid decisions about interventions and supports for struggling ELs.

### *Limitations*

The data used in this study were generated from EL students attending ORRF schools (Baker et al., 2011). Decision thresholds presented here may not generalize to all children in all schools. Our sample included a wide range of ELs, with small pockets of students (1%) speaking multiple languages other than Spanish. As a result, findings may differ for ELs in different types of programs, from different language groups, or for an entirely Spanish-speaking sample. Nevertheless, the use of sensitivity as a metric to set decision thresholds does not depend on base rates, which, unlike predictive values, should minimize differences across any schools that aim to achieve the same criterion level of performance. Indeed, the difference between cut scores for ELs and EPs, when based on predictive values, differed tremendously—by 39 letter sounds for NWF in the spring of Grade 1 (Smolkowski, Baker, & Seeley, 2008)—the present study shows a difference of only four letter sounds.

### *Implications for Practice*

D6 decision thresholds yield similar predicative utility for students learning English and children who are native English speakers. Schools may select to use the cut points recommended by Smolkowski and Cummings (2016) or adopt the thresholds described in the present study; both provide accurate and sensitive estimates of future reading performance for both groups. While the same thresholds may hold adequate value for prediction of reading disability with EL and native English-speaking students, associated causal factors likely vary considerably for these two groups. Consequently, we encourage schools to carefully account for group and individual differences in planning, delivering, and monitoring instructional interventions for ELs and their native English-speaking peers identified at risk for reading difficulties using these screening measures.

### Authors' Note

The data used in this report were previously published in Baker et al. (2012); Baker et al. (2008); Baker et al. (2011); Fien et al. (2008); and Smolkowski and Cummings (2016).

### Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

## Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: The research was supported by an Oregon Reading First grant from the U.S. Department of Education (S357A0020038). This research was also supported by two grants from the Institute of Education Sciences, U.S. Department of Education (R305A150325, R324A090104). The opinions expressed are those of the authors and do not represent views of Oregon Research Institute, Southern Methodist University, the University of Oregon, the University of Maryland, or the U.S. Department of Education.

## ORCID iDs

Kelli D. Cummings https://orcid.org/0000-0002-3703-9852
Keith Smolkowski https://orcid.org/0000-0003-2565-3297

## Supplemental Material

Supplemental material for this article is available on the *LDQ* along with the online version of this article.

## Notes

1. Technically, as sensitivity increases, specificity decreases *or remains the same*, and vice versa.
2. Scheffel et al. (2012) incorrectly calculated sensitivity and specificity, discussed below, but the points about the trade-offs still hold even for their miscalculated sensitivity and specificity values.
3. The assumptions for the two-proportions test, used for area under the ROC curve (AUC) comparisons in Hosp et al. (2011), may have been invalid. The test assumes proportions have values in the unit interval [0, 1], whereas the AUC uses half the range [0, .5], so the standard error estimate may be incorrect. The authors did not report an adjustment for the AUC.

## References

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

Baker, D. L., Park, Y., Baker, S. K., Basaraba, D. L., Kame'enui, E. J., & Thomas Beck, C. (2012). Effects of a paired bilingual reading program and an English-only program on the reading performance of English learners in Grades 1–3. *Journal of School Psychology*, *50*, 737-758.

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*, 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*, 307–331. doi:10.1086/661995

Betts, J., Reschly, A., Pickart, M., Heistad, D., Sheran, C., & Marston, D. (2008). An examination of predictive bias for second grade reading outcomes from measures of early literacy skills in kindergarten with respect to English-language learners and ethnic subgroups. *School Psychology Quarterly*, *23*, 553–570.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum.

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*, 284–295.

Cummings, K. D., & Smolkowski, K. (2015). Bridging the gap: Selecting students at risk for academic difficulties. *Assessment for Effective Intervention*, *44*, 55–61. doi:10.1177/1534508415590396

de Ramírez, R. D., & Shapiro, E. S. (2006). Curriculum-based measurement and the evaluation of reading skills of Spanish-speaking English language learners in bilingual education classrooms. *School Psychology Review*, *35*, 356–369.

Dynamic Measurement Group. (2008). *DIBELS 6th Edition technical adequacy information* (Technical Report No 6). Eugene, OR: Author. Retrieved from http://dibels.org/pubs.html

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*, 391–408.

García, E. (2015). *Inequality at the starting gate: Cognitive and noncognitive skills gaps between 2010–2011 kindergarten classmates* (Report). Washington, DC: Economic Policy Institute.

Gerber, M., Jimenez, T., Leafstedt, J., Villaruz, J., Richards, C., & English, J. (2004). English reading effects of small-group intensive intervention in Spanish for K–1 English learners. *Learning Disabilities Research & Practice*, *19*, 239–251.

Good, R. H., III, 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 IV* (pp. 679–700). Washington, DC: National Association of School Psychologists.

Good, R. H., III, & Kaminski, R. (2002). *Dynamic Indicators of Basic Early Literacy Skills™ 6th edition: Administration and scoring guide*. Eugene: University of Oregon Institute for the Development of Educational Achievement.

Good, R. H., III, Kaminski, R. A., Cummings, K., Dufour-Martel, C., Peterson, K., Powell-Smith, K., . . . Wallin, J. (2011). *DIBELS Next assessment manual*. Eugene, OR: Dynamic Measurement Group.

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: University of Oregon, Center on Teaching and Learning.

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*, 257–288.

Greenland, S., Senn, S. J., Rothman, K. J., Carlin, J. B., Poole, C., Goodman, S. N., & Altman, D. G. (2016). Statistical tests, *p*-values, confidence intervals, and power: A guide to misinterpretations. *The American Statistician*, *70*(2), 1–12.

Harcourt Educational Measurement. (2002). *Stanford Achievement Test*. San Antonio, TX: Author.

Honig, B., Diamond, L., & Gutlohn, L. (2000). *Teaching reading sourcebook for kindergarten through eighth grade*. Novato, CA: Arena Press.

Hosp, J. L., Hosp, M. A., & Dole, J. K. (2011). Potential bias in predictive validity of universal screening measures across disaggregation subgroups. *School Psychology Review*, *40*, 108–131.

Jacowitz, K. E., & Kahneman, D. (1995). Measures of anchoring in estimation tasks. *Personality and Social Psychology Bulletin*, *21*, 1161–1166.

Johnson, E. S., Jenkins, J. R., Petscher, Y., & Catts, H. W. (2009). How can we improve the accuracy of screening instruments? *Learning Disabilities Research and Practice*, *24*, 174–185. doi:10.1111/j.1540-5826.2009.00291.x

Kim, J. S., Vanderwood, M. L., & Lee, C. Y. (2016). Predictive validity of curriculum-based measures for English learners at varying English proficiency levels. *Educational Assessment*, *21*(1), 1–18.

Klingner, J. K., Artiles, A. J., & Méndez Barletta, L. (2006). English language learners who struggle with reading: Language acquisition or learning disabilities? *Journal of Learning Disabilities*, *39*, 108–128.

Malhotra, R., & Indrayan, A. A. (2010). A simple nomogram for sample size for estimating sensitivity and specificity of medical tests. *Indian Journal of Ophthalmology*, *58*, 519–522. doi:10.4103/0301-4738.71699

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

Meehl, P. E. (1971). High school yearbooks: A reply to Schwarz. *Journal of Abnormal Psychology*, 77, 143–148.

Nakamoto, J., Lindsey, K. A., & Manis, F. R. (2012). Development of reading skills from K-3 in Spanish-speaking English language learners following three programs of instruction. *Reading and Writing: An Interdisciplinary Journal*, *25*, 537–567. doi:10.1007/s11145-010-9285-4

National Center on Response to Intervention. (2010). *Essential components of RTI—A closer look at response to intervention*. Washington, DC: U.S. Department of Education, Office of Special Education Programs, National Center on Response to Intervention.

Northwest Evaluation Association. (2004). *Reliability and validity estimates: NWEA Achievement Level tests and Measures of Academic Progress*. Lake Oswego, OR: Author.

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, NY: Oxford.

Rice, M. E., & Harris, G. T. (2005). Comparing effect sizes in follow-up studies: ROC area, Cohen's d, and r. *Law and Human Behavior*, *29*, 615–620.

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

Rueda, R., & Windmueller, M. P. (2006). English language learners, LD, and overrepresentation: A multiple-level analysis. *Journal of Learning Disabilities*, *39*, 99–107.

Sandberg, K. L., & Reschly, A. L. (2011). English learners: Challenges in assessment and the promise of Curriculum-Based Measurement. *Remedial and Special Education*, *32*, 144–154. doi:10.1177/0741932510361260

SAS Institute. (2016). *SAS/STAT® 14.2 user's guide*. Cary, NC: Author.

Scheffel, D., Lefly, D., & Houser, J. (2012). The predictive utility of DIBELS reading assessment for reading comprehension among third-grade English language learners and English speaking children. *Reading Improvement*, *49*, 75–92.

Shapiro, E. S., Solari, E., & Petscher, Y. (2008). Use of a measure of reading comprehension to enhance prediction on the state high stakes assessment. *Learning and Individual Differences*, *18*, 316–328. doi:10.1016/j.lindif.2008.03.002

Smolkowski, K., Baker, S., & Seeley, J. (2008, February). ROC done right, part 1: The statistical evaluation of diagnostic tests. Paper presented at the Pacific Coast Research Conference, San Diego, CA.

Smolkowski, K., & Cummings, K. D. (2015). Evaluation of diagnostic systems: The selection of students at risk of academic difficulties. *Assessment for Effective Intervention*, *41*, 41–54. doi:10.1177/1534508415590386

Smolkowski, K., & Cummings, K. D. (2016). Evaluation of the DIBELS (Six Edition) diagnostic system for the selection of native and proficient English speakers at risk for reading difficulties. *Journal of Psychoeducational Assessment*, *34*, 103–118. doi:10.1177/0734282915589017

Smolkowski, K., Cummings, K. D., & Stryker, L. (2016). An introduction to the statistical evaluation of fluency measures with signal detection theory. In K. D. Cummings & Y. Petscher (Eds.), *The fluency construct: Curriculum-based measurement concepts and applications*. New York, NY: Springer.

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

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

U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics. (2016). *Digest of Education Statistics 2015, Table 204.20*. Washington, DC: Author.

Vanderwood, M. L., Tung, C. Y., & Checca, C. J. (2014). Predictive validity and accuracy of oral reading fluency for English learners. *Journal of Psychoeducational Assessment*, *32*, 249–258.