School Psychology Review,
2011, Volume 40, No. 1, pp. 108–131

Potential Bias in Predictive Validity of Universal Screening Measures Across Disaggregation Subgroups

John L. Hosp
University of Iowa

Michelle A. Hosp
Iowa Department of Education

Janice K. Dole
University of Utah

Abstract. Universal screening measures are an integral component of any tiered system of instructional delivery. Recent studies of screening measures have often excluded examinations of bias in predictive validity. The present study examined a common screening instrument for evidence of bias in predictive validity across the four disaggregation categories of the No Child Left Behind Act. Performance of 3,805 students in Grades 1–3 on the Nonsense Word Fluency and Oral Reading Fluency measures of the Dynamic Indicators of Basic Early Literacy Skills were examined cross-sectionally in relation to a state criterion-referenced test. Bias in predictive validity was found, but varied by grade and by disaggregation category. Implications are discussed.

Universal screening is a crucial component of any comprehensive system of assessment (Salvia, Ysseldyke, & Bolt, 2009), especially those used within a problem-solving or response to intervention framework (Batsche et al., 2005). Efficient and effective delivery of the most appropriate interventions to the right students requires a consistent and accurate process of identifying which students need what help (Hosp & Ardoin, 2008). While providing this help, it is also crucial to be able to accurately and efficiently judge each student's response (Barnett et al., 2007). Universal screening involves the assessment of all students within a classroom, grade, school, or district on measures that are valid indicators of important academic or social/emotional outcomes (Ikeda, Neessen, & Witt, 2008). These assessments should be quick to administer, score, and provide information that leads to valid inferences about those outcomes (Hosp & Ardoin, 2008). These inferences are the decisions that need to be made in identifying each student's level of need as well as grouping students with similar needs.

Given the recent focus on universal screening that has come from a renewed emphasis on problem solving in delivering educational services, it is no surprise that there have been recent advances in the development

Correspondence regarding this article should be addressed to John L. Hosp, College of Education, N264 Lindquist Center, Iowa City, IA 52242; e-mail: john-hosp@uiowa.edu

of screening measures (Catts, Fey, Zhang, & Tomblin, 2001; Foorman, Francis, Fletcher, Schatschneider, & Mehta, 1998; O'Connor & Jenkins, 1999). However, there has also been increased scrutiny to ensure that they result in reliable data that provide accurate classification of students as needing intervention or not. Ritchey and Speece (2004) explored characteristics of screening assessment that should be considered in the early identification of reading disabilities. Differences in skills measured, performance tasks required, and content coverage are all characteristics that can affect the classification accuracy of a measure. The timing of a measure is important in terms of both the interval between screening measurement and outcome measurement as well as when the screening measurement takes place in a developmental sequence. Selection of outcome is also important as prediction of more proximal outcomes is likely to be more accurate than prediction of more distal ones. Jenkins, Hudson, and Johnson (2007) suggested additional factors to consider when developing and using screening measures such as accounting for the severity of the problem, different levels of risk (use of the dichotomous at risk/not at risk or a polytomous system), and the inclusion of cross-validation of screening measures that is a crucial measurement component to the development of any measure (Haladyna, 2006).

Predictive Validity

A key component in the determination of the quality of a screening measure is its predictive validity. Predictive validity is an indication of how well performance on a criterion measure is predicted by performance on a screening measure when there is a difference in the time of administration (typically 3–5 months) between the two measures (Salvia et al., 2009). The criterion measure is also described as a meaningful outcome (Ikeda et al., 2008) such as performance on the state high-stakes test.

Researchers have evaluated the predictive validity of Nonsense Word Fluency (NWF) and Oral Reading Fluency (ORF) as compared to norm-referenced tests of reading (e.g., Woodcock Reading Mastery Test; Woodcock, 1998; see Ritchey, 2008) as well as the mandated high-stakes state tests of many states (e.g., Florida—Buck & Torgesen, 2002; Washington—Stage & Jacobson, 2001). Predictive validity coefficients for both NWF and ORF typically average between .65 and .75 (cf. Hintze & Silberglitt, 2005; Ritchey, 2008; Roehrig, Petscher, Nettles, Hudson, & Torgesen, 2007; Shanahan, 2003).

Catts, Petscher, Schatschneider, Bridges, and Mendoza (2009) recently extended this predictive validity work by looking for floor effects that might influence the predictive accuracy of screening measures, particularly Initial Sound Fluency, Phoneme Segmentation Fluency, NWF, and ORF from the Dynamic Indicators of Basic Early Literacy Skills (DIBELS; Good et al., 2004). Floor effects occur when the lower end of the performance range for a scale does not go low enough to adequately describe participants' performance (Drew, Hardman, & Hosp, 2008). This is demonstrated by a large number of individuals receiving scores near the minimum possible score such that the scores for a group are "bunched" near the minimum performance. This can have a negative effect on predictive validity because of the restriction of range of participants' performance. Using nearly 19,000 students from Florida, Catts et al. (2009) demonstrated that floor effects were present, which reduced the predictive validity of the data. The effect on predictive validity was most pronounced in kindergarten and Grade 1, with decreasing effect in Grade 2 and little to no effect in Grade 3. Catts et al. conclude that more sensitive measures of early literacy are needed to overcome these floor effects and the effect they have on predictive validity.

The Importance of Examining Bias in Predictive Validity

The studies mentioned above have contributed to our understanding of screening measures and some of the potential issues that affect the results provided and the inferences made from those results. However, they did

not examine the potential differential performance of screening measures across different subgroups of students. The achievement gap between subgroups of students has been a longstanding and persistent issue in American education (Rampsey, Dion, & Donahue, 2009). This is one reason that disaggregation across various traditionally underperforming subgroups was required for states to demonstrate adequate yearly progress under the No Child Left Behind Act (NCLB; 2002). By explicitly disaggregating the performance of various subgroups (i.e., students from economically disadvantaged backgrounds, students with limited English proficiency, students with disabilities [SwD], and students from various racial/ethnic backgrounds), schools, districts, and states would be able to determine whether they were meeting the needs of all groups of students. However, this analysis is on state-identified outcome measures only and does not address the potential for differential prediction on screening measures.

Although overall classification accuracy is an important consideration when evaluating screening measures, bias in predictive validity is also an important consideration (Cole & Moss, 1993). Bias in predictive validity (also referred to as "differential prediction") is a difference in the quality of inferences when making a judgment of individuals from one group rather than another (Helms, 2006). That is, it is a difference between two groups in the predictive validity of a measure. Instruments used for universal screening are often characterized by high rates of under- or overidentification. This is part of the effect that test use has on students that has been shown to differentially affect different subgroups of students (Cleary, Humphreys, Kendrick, & Wesman, 1975). It is also often implicated in the disproportionate representation of minority students in special education (Hosp & Reschly, 2003) and differential provision of services, in that not only might some individuals, but some groups, might be overidentified yet underserved (Donovan & Cross, 2002). With the increased emphasis on assessment and accountability as mandated through NCLB as well as Race to the Top (2010) and the increased alignment of the Individuals with Disabilities Education Act (2004) with the proposed revisions to the Elementary and Secondary Education Act (NCLB, 2002), the influence of assessment on students and the importance of examining potential bias in predictive validity arguably has never been higher.

Unfortunately, bias in predictive validity is something that is evaluated less frequently than it should (Betts et al., 2008). For example, The National Center on Response to Intervention conducts technical reviews of screening instruments for reading and math. To date, the technical review committee has reviewed nine reading-related screening measures, and only two (DIBELS ORF and STAR Reading [Renaissance Learning, 2011]) have provided evidence of predictive validity across disaggregation groups (see http://www.rti4success.org).

There have been a few studies to examine differential predictive validity (i.e., with a span of >3 months between predictor and criterion) of screening measures across disaggregation groups. Wiley and Deno (2005) found differences between English learners (EL) and English fluent (ES) students on both Maze and ORF tasks as compared to a state high-stakes test at Grades 3 and 5. Roehrig et al. (2007) found no differences for students receiving free or reduced-price lunch, EL students, or African American and Hispanic third-grade students in prediction of the state high-stakes test using ORF. Betts et al. (2008) found no difference for EL students, African American students, or Asian-American students, but some difference between White and Latino students when predicting second-grade reading outcomes from kindergarten screening assessments. Last, Fien et al. (2008) found that 7 of 24 comparisons between EL and ES students demonstrated differential prediction, but that 5 of the 7 were from winter kindergarten assessments, which was consistent with Catts et al.'s (2009) concern about floor effects and suggesting that the measures may differentially affect different subgroups of students. When examined as a group, no clear pattern of differential prediction has been consistent across groups, grade levels, or samples.

Purpose of the Study

Given the importance of screening in response to intervention (Hughes & Dexter, 2007), the concerns detailed by Catts et al. (2009), and the need to demonstrate accountability for disaggregated subgroups within NCLB, the purpose of the current study was to examine the possibility of bias in predictive validity (including differential prediction and differential floor effects) across the disaggregation categories included in NCLB. As such, the research questions guiding this study were as follows:

1. How well do benchmark scores on the NWF and ORF measures of the DIBELS predict grades 1–3 scores on a state criterion-referenced test when examined across the disaggregation categories of NCLB?
2. How much does the accuracy of prediction of the NWF and ORF measures of the DIBELS on a state criterion-referenced test vary as a function of level of performance when examined across the disaggregation categories of NCLB?

These research questions served as the basis for the following hypotheses:

H_1. Benchmark scores on the NWF and ORF measures of the DIBELS will differentially predict scores on a state criterion-referenced test when examined across the disaggregation categories of NCLB.

H_2. Accuracy of prediction of the NWF and ORF measures of the DIBELS on a state criterion-referenced test will vary as a function of level of performance when examined across the disaggregation categories of NCLB.

Method

Participants

Participants were 3,805 students enrolled in Grades 1–3 of Utah's Reading First schools during the 2006–2007 school year. This sample included all the students in Utah's Reading First schools who had data on both measures. The entire sample of students was 50.8% male, 71.8% eligible for free or reduced-price lunch, 25.3% EL, 9.4% students with disabilities, 45.8% White, 38.7% Hispanic, 8.7% American Indian, 2.6% Pacific Islander, 2.1% African American, and 1.1% Asian. See Table 1 for the demographic characteristics broken out by each subgroup used in the analysis for each grade level. Analyses indicated no differences between the demographic profile of the final sample and overall school demographics.

Measures

As part of Utah's Reading First, all children were required to be administered a screening instrument at least three times per year in order to predict which students were likely to not reach proficiency on the state's criterion-referenced test, which is used to report adequate yearly progress to the U.S. Department of Education as a condition of NCLB. The reading coaches, reading coordinators, and administrators from the participating schools chose to use the DIBELS as their screening measures, which was then implemented in all Reading First schools in Utah. For the purposes of this study, only NWF and ORF were included because these are the only DIBELS measures administered in Grades 1–3, which are grades in which an outcome measure is also administered.

NWF. This is a standardized, individually administered measure of a student's ability to use letter–sound correspondence to decode short consonant–vowel–consonant (CVC) and vowel–consonant (VC) nonsense words. Given a page of these words, the student must verbally produce either the individual letter sounds or each nonsense word. The student's score is the number of correct letter–sound correspondences produced within 1 min. Reliability for NWF with first-grade students has been reported as .94 for test–retest (Harn, Stoolmiller, & Chard, 2008) and .83 (Mdn = .67 to .88 range) for 1-month alternate form (Good et al., 2004).

Table 1
Demographic Characteristics of the Participants in Each Subgroup at Each Grade Level

Group	*n*	Male	FRL	EL	SwD	AA	AI	As	W	H	PI	O
				Grade 1 (*n* = 1353)								
FRL	945	483	—	368	70	19	78	9	311	487	33	8
non-FRL	408	213	—	35	30	7	16	2	311	59	10	2
EL	403	203	368	—	24	7	30	6	9	337	10	4
non-EL	950	494	577	—	77	19	64	5	613	210	33	6
SwD	101	73	70	24	—	4	6	0	56	31	2	2
non-SwD	1252	624	875	379	—	22	88	11	566	516	41	8
AI	94	39	78	30	6	—	—	—	—	—	—	—
W	622	324	311	9	56	—	—	—	—	—	—	—
H	547	286	487	337	31	—	—	—	—	—	—	—
				Grade 2 (*n* = 1241)								
FRL	886	459	—	286	73	23	48	11	339	428	28	9
non-FRL	351	187	—	22	33	3	18	3	279	43	4	1
EL	311	150	286	—	17	6	3	8	7	276	10	1
non-EL	930	497	600	—	89	21	63	6	611	197	23	9
SwD	106	72	73	17	—	1	10	1	60	32	1	1
non-SwD	1135	575	813	294	—	26	56	13	558	441	32	9
AI	66	32	48	3	10	—	—	—	—	—	—	—
W	618	340	339	7	60	—	—	—	—	—	—	—
H	473	234	428	276	32	—	—	—	—	—	—	—
				Grade 3 (*n* = 1088)								
FRL	766	397	—	221	107	22	40	10	268	401	18	7
non-FRL	322	155	—	26	34	7	20	8	234	49	3	1
EL	247	139	221	—	33	8	0	9	8	211	7	4
non-EL	841	413	545	—	108	21	60	9	494	239	14	4
SwD	141	91	107	33	—	2	8	3	72	55	1	0
non-SwD	947	461	659	214	—	27	52	15	430	395	20	8
AI	60	28	40	0	8	—	—	—	—	—	—	—
W	502	244	268	8	72	—	—	—	—	—	—	—
H	450	244	401	211	55	—	—	—	—	—	—	—

Note. FRL = students receiving free/reduced-price lunch; non-FRL = students not receiving free/reduced-price lunch; EL = English learners; non-EL = English-proficient students; SwD = students with disabilities; non-SwD = students without IEPs; AA = African American; AI = American Indian; As = Asian; W = White; H = Hispanic; PI = Pacific Islander; O = other race/ethnicity.

ORF. This is a standardized, individually administered measure of the accuracy and rate of a student's ability to orally read connected text. Given a grade-level passage of previously unseen material, the student reads aloud for 1 min. The number of words read correctly in that minute is recorded. Three separate passages are administered with the student's median words read correctly score serving as the student's recorded score. Reliability of ORF has been reported as .95 for alternate form (Good, Kaminski, Smith, & Bratten, 2001) and .96 for test–retest (Catts et al., 2009).

Utah State Criterion-Referenced Tests (UCRTs). The UCRTs are group-administered tests given to all students in Grades 1–8 in the spring of each school year. The questions are in multiple-choice format with

students recording their answers on a computerized Scantron sheet. The items are aligned with the state core curriculum with cut scores established by the Utah State Office of Education to determine the minimum score a student must receive to achieve the level of proficiency in the state curriculum. Because NWF and ORF are designed to predict reading outcomes, the English/Language Arts component was used as the outcome in this study. Reliability of the UCRTs was reported as .92 (Kuder-Richardson 20) and .93 (stratified alpha) for internal consistency; criterion validity was reported as .65 with the Grade 3 Iowa Test of Basic Skills (Hoover et al., 2003; Utah State Office of Education, 2007). Analyses also suggest that the UCRT meets standards as a nonbiased instrument (Utah State Office of Education, 2007).

Procedures

All measures were administered by classroom teachers or reading coaches. Training for administration and scoring of the DIBELS measures was conducted by outside expert consultants hired by the Utah Reading First Director to conduct multiple two-day trainings for educators across the state. District-based coaches and coordinators were also trained in using the DIBELS administration integrity checklists to use while observing practice administrations of the measures. Although data from those checklists are not available for analysis in this study, no individual was allowed to administer the measures without demonstrating administration and scorning accuracy $>95\%$ to a trainer.

The Utah Reading First evaluation team provided a schedule for DIBELS administration to all participating schools in order to have consistency in administration times across schools. Two-week windows for administration were provided for fall (2– 4 weeks after the first day of school, typically occurring in early September), winter (2 weeks in January equidistant from the beginning and end of the school year), and spring (2–4 weeks before the last day of school, typically occurring in early May). NWF was administered at all three time points in Grade 1 whereas ORF was administered for winter and spring in Grade 1 and all three time points in Grades 2 and 3. The UCRTs are administered over a 2-week period typically occurring in April or early May. The spring DIBELS window was scheduled so as not to overlap with the UCRT administration window to reduce scheduling burden.

Data Analysis

To make comparisons, four sets of analyses were conducted. These aligned with the disaggregation categories as required through NCLB: economic disadvantage (operationalized as receiving free/reduced-price lunch or not), limited English proficiency (identification as an English learner or not), disability status (identification as having a disability or not), and race/ethnicity (White, Hispanic, and American Indian—these three groups with large enough samples for analysis). The sample sizes for each analysis can be found in Table 2.

Two methods of analysis were used to address the research questions for this study. First, Receiver Operating Characteristic (ROC) curves were calculated for each group within each disaggregation category, for each measure, at each grade level. ROC curve analysis is a method of judging the diagnostic efficiency of a measure (Swets, 1996). The three indexes included from the ROC curve analyses are sensitivity (SE; i.e., the proportion of students correctly classified as nonproficient on both measures being compared), specificity (SP; i.e., the proportion of students correctly classified as proficient on both measures), and area under the curve (AUC). AUC is a probability ranging from 0.5 to 1.0 that provides the probability of a predictor correctly classifying a pair of students from two different categories (e.g., proficient, nonproficient), and can be used as an effect size statistic (Swets, 1988). The SE, SP, and AUC were all compared between groups within disaggregation categories using a two-proportions test (Sprinthall, 2003). Determination for significance was adjusted for multiple comparisons

with $p < .001$ being used given 33 comparisons in each family of analyses (Drew et al., 2008).

The second method of analysis was the use of quantile regression. Quantile regression is similar to ordinary least-squares regression in that it attempts to minimize the sum of squared residuals, but rather than calculating a *line* to represent the best-fit data, quantile regression can provide best-fit points by asymmetrically weighting the data above and below the point of interest (Koenker, 2005). These estimates can be plotted in a simple line graph to illustrate the change in correlation between two variables across levels of the predictor variable. By plotting the quantile regression lines for multiple groups on a single graph, the differential effect of the relation between the two variables for different disaggregation groups can be examined as well as the presence of floor and ceiling effects.

Results

Descriptive Statistics

The means and standards deviations of the performance of each disaggregated group are shown in Table 2. Although on average the traditionally underrepresented groups (FRL, EL, SwD, Hispanic, and American Indian) appeared to perform below their comparison group (non-FRL, non-EL, non-SwD, and White, respectively), this was not a hypothesis of the current study and therefore not tested for statistical significance. Also, all groups showed within-grade growth on both NWF and ORF, but again these gains were not compared to a norm or criterion standard to test for their significance. Within-grade growth cannot be compared for the UCRTs as it is only administered once per year. Cross-grade comparisons are not made here because this sample is cross-sectional, rather than longitudinal.

ROC Curve Analyses

To answer the first research question, "How well do benchmark scores on the NWF and ORF measures of the DIBELS predict Grade 1–3 scores on a state criterion-referenced test when examined across the disaggregation categories of NCLB?," a series of ROC curves were calculated. The AUC index was evaluated as good if it was >.80 (Metz, 1978); the SE index was evaluated as good if it was >.80 (Carran & Scott, 1992); the SP index was also evaluated as good if it was >.80. In addition, the AUC, SE, and SP indexes between the groups were compared using a two-proportions test.

Economic disadvantage. Results of the ROC curve analyses for the economic disadvantage disaggregation comparisons are shown in Table 3. Overall, the AUCs fell into the desired range for screening measures. All AUCs for both groups, except for the three administrations of the NWF measure for the FRL group, were >.80. For SE, ORF in Grades 2 and 3 exceeded the .80 criterion for both the FRL and non-FRL groups with the exception of spring Grade 3 for the non-FRL group (.79). The only measurement for either group to meet the criterion for Grade 1 was winter ORF for the non-FRL group. The opposite was true for SP; no measurements in Grades 2 or 3 met the criterion, but all except winter NWF and winter ORF for the FRL group did in Grade 1. Using the conservative criterion of $p < .001$, two measurements demonstrated differences in AUC: two in SE, and three in SP. Most of these differences were in Grade 1, with two of the SP differences in ORF for Grade 2. There were no significant differences in AUC or SE for Grade 2, and no significant differences for any index at Grade 3.

Limited English proficiency. Results of the ROC curve analysis for the limited English proficiency disaggregation comparisons are shown in Table 4. Overall, the AUCs fell into the desired range for screening measures for ORF but not for NWF. For SE, ORF in Grades 2 and 3 exceeded the .80 criterion for both groups. In Grade 1, only the winter ORF measurement for the EL group exceeded the criterion. The opposite was true for SP; no measurements in Grades 2 or 3 met the criterion, but all except winter NWF and winter

Table 2
Means and Standard Deviations for Each Group in Each Grade

				Fall		Winter		Spring	
Measure	Grade	Group	*n*	Mean	*SD*	Mean	*SD*	Mean	*SD*
NWF	1	FRL	945	34.2	22.5	61.3	27.5	78.3	36.0
		non-FRL	408	40.4	26.2	69.7	32.8	86.1	36.8
		EL	403	30.4	20.5	55.4	25.6	74.0	34.5
		non-EL	950	38.2	24.6	67.0	30.1	82.4	35.4
		SwD	101	21.4	17.5	45.6	24.5	59.2	35.5
		non-SwD	1252	37.2	23.9	65.2	29.3	82.3	35.9
		W	622	39.0	25.8	68.1	31.0	82.9	35.7
		H	547	31.1	19.7	56.8	25.1	73.5	32.3
		AI	94	43.6	24.0	74.3	28.7	101.0	47.1
ORF	1	FRL	945	—	—	27.9	25.9	49.0	30.3
		non-FRL	408	—	—	40.6	34.0	65.1	35.7
		EL	403	—	—	21.6	20.9	41.7	27.4
		non-EL	950	—	—	35.6	31.0	58.8	33.7
		SwD	101	—	—	16.6	18.0	33.5	27.6
		non-SwD	1252	—	—	32.8	29.5	55.4	32.7
		W	622	—	—	39.0	33.3	62.0	35.0
		H	547	—	—	22.0	20.1	44.4	27.7
		AI	94	—	—	38.2	28.0	53.3	29.4
	2	FRL	886	43.4	28.9	66.0	34.9	79.4	37.1
		non-FRL	351	56.8	33.8	82.7	38.5	96.2	38.5
		EL	311	36.0	28.2	56.9	35.9	70.3	38.6
		non-EL	930	50.6	31.1	75.5	35.1	88.9	37.4
		SwD	106	22.4	20.6	39.2	32.1	50.7	34.4
		non-SwD	1135	49.3	30.7	73.6	35.6	87.2	37.1
		W	618	52.4	32.8	78.3	37.4	91.0	38.6
		H	473	40.0	27.4	61.7	34.2	75.7	36.4
		AI	66	44.8	27.8	62.6	31.7	77.5	33.2
	3	FRL	766	69.5	34.2	82.4	37.2	99.1	37.9
		non-FRL	322	82.3	35.8	95.5	37.4	111.4	36.0
		EL	247	56.0	32.0	69.0	36.8	86.0	37.3
		non-EL	841	79.0	34.6	92.0	36.6	108.0	36.5
		SwD	141	42.6	32.5	51.5	37.7	66.9	40.9
		non-SwD	947	77.7	33.1	91.0	35.0	107.6	34.4
		W	502	81.6	35.5	94.4	37.6	110.3	37.2
		H	450	66.1	32.9	78.8	35.8	96.3	36.5
		AI	60	65.2	31.1	77.8	34.9	95.6	37.7
UCRT[a]	1	FRL	945	—	—	—	—	159.7	11.2
		non-FRL	408	—	—	—	—	167.3	13.6
		EL	403	—	—	—	—	155.6	9.9
		non-EL	950	—	—	—	—	164.5	12.5
		SwD	101	—	—	—	—	154.4	11.6
		non-SwD	1252	—	—	—	—	162.5	12.3
		W	622	—	—	—	—	166.9	12.8
		H	547	—	—	—	—	157.3	10.2
		AI	94	—	—	—	—	158.6	10.2

(Table 2 continues)

Table 2 Continued
Means and Standard Deviations for Each Group in Each Grade

				Fall		Winter		Spring	
Measure	Grade	Group	*n*	Mean	*SD*	Mean	*SD*	Mean	*SD*
	2	FRL	886	—	—	—	—	160.1	10.6
		non-FRL	351	—	—	—	—	166.2	10.6
		EL	311	—	—	—	—	156.0	10.7
		non-EL	930	—	—	—	—	163.8	10.5
		SwD	106	—	—	—	—	154.5	9.4
		non-SwD	1135	—	—	—	—	162.5	10.8
		W	618	—	—	—	—	165.2	10.9
		H	473	—	—	—	—	158.0	10.2
		AI	66	—	—	—	—	159.3	8.2
	3	FRL	766	—	—	—	—	160.9	9.8
		non-FRL	322	—	—	—	—	164.5	9.9
		EL	247	—	—	—	—	156.7	9.9
		non-EL	841	—	—	—	—	163.6	9.6
		SwD	141	—	—	—	—	155.4	9.4
		non-SwD	947	—	—	—	—	162.8	9.7
		W	502	—	—	—	—	164.7	9.4
		H	450	—	—	—	—	159.4	10.1
		AI	60	—	—	—	—	158.8	7.2

Note. NWF = Nonsense Word Fluency; ORF = Oral Reading Fluency; UCRT = Utah Criterion-Referenced Test; FRL = students receiving free/reduced-price lunch; non-FRL = students not receiving free/reduced-price lunch; EL = English learners; non-EL = English proficient students; SwD = students with disabilities; non-SwD = students without IEPs; W = White; H = Hispanic; AI = American Indian. Mean for NWF is the mean number of correct letter sounds per minute. For ORF, the mean is the mean number of words read correctly in one minute.
[a]The UCRT score reported here is the scaled score. This scale is equated across grade levels so that a score of 160 always indicates performance at the median for that grade.

ORF for the EL group did in Grade 1. Using the conservative criterion of $p < .001$, no comparisons between the EL and non-EL groups demonstrated significant differences. In Grade 1, the winter NWF and winter ORF comparisons were significant for SP. Winter ORF was also significant for SE. In Grade 2, the winter and spring ORF comparisons were significant for SP, but not for SE. In Grade 3, all SE and SP comparisons were significant.

Disability status. Results of the ROC curve analysis for the disability status disaggregation comparisons are shown in Table 5. Overall, the ORF AUCs fell into the desired range for screening measures, but the NWF AUCs did not. The one exception is Grade 1 winter ORF AUC for the SwD group (.794). For SE, ORF met the criterion at all measurement points for the SwD group, but only for Grades 2 and 3 (with the exception of winter ORF at Grade 2, .794) for the non-SwD group. SP again had nearly the opposite pattern with the non-SwD group measurements at Grade 1, except for winter ORF (.795) meeting the criterion. For the SwD group, only spring NWF met the criterion. None of the AUC comparisons met the criterion for significance; however, almost all of the SE and SP comparisons met the significance criterion with the exception of winter ORF (SE $p = .033$) and spring NWF (SP $p = .010$).

Race/ethnicity. Results of the ROC curve analysis for the race/ethnicity disaggregation comparisons are shown in Tables 6 and

Table 3
ROC Two-Proportions Test Results for the Economic Disadvantage Disaggregation Analysis

		FRL							non-FRL						*p*		
Grade	Measure	BM	SE	SP	AUC	PPP	NPP	K	SE	SP	AUC	PPP	NPP	K	SE	SP	AUC
1	F NWF	24	.535	.853	.761	.764	.665	.384	.618	.875	.845	.648	.860	.501	.001	.290	.001
	W NWF	50	.530	.790	.733	.692	.648	.319	.600	.862	.817	.617	.853	.466	.020	.002	.001
	S NWF	50	.324	.917	.741	.775	.598	.245	.373	.936	.802	.683	.801	.360	.075	.230	.018
	W ORF	20	.769	.754	.849	.739	.784	.523	.809	.838	.901	.650	.922	.601	.089	.001	.011
	S ORF	40	.732	.851	.867	.816	.778	.586	.627	.899	.884	.697	.867	.543	.000	.018	.390
2	F ORF	44	.845	.671	.862	.666	.868	.517	.812	.737	.892	.496	.925	.451	.134	.014	.121
	W ORF	68	.821	.721	.874	.692	.844	.534	.812	.820	.888	.583	.924	.543	.689	.000	.453
	S ORF	90	.870	.663	.866	.660	.865	.507	.859	.771	.889	.545	.945	.526	.575	.000	.230
3	F ORF	77	.874	.595	.843	.591	.875	.439	.815	.693	.822	.468	.918	.406	.019	.003	.384
	W ORF	92	.865	.577	.841	.582	.875	.426	.827	.656	.840	.443	.919	.373	.121	.017	.968
	S ORF	110	.865	.613	.846	.606	.881	.464	.790	.668	.827	.441	.904	.361	.004	.089	.430

Note. ROC = receiver operating characteristic; FRL = students receiving free/reduced-price lunch; BM = benchmark score; SE = sensitivity; SP = specificity; AUC = area under the curve; PPP = positive predictive power; NPP = negative predictive power; K = kappa; F = fall; W = winter; S = spring; NWF = Nonsense Word Fluency; ORF = Oral Reading Fluency.

Table 4
ROC Two-Proportions Test Results for the Limited English Proficiency Disaggregation Analysis

		EL							non-EL						*p*		
Grade	Measure	BM	SE	SP	AUC	PPP	NPP	K	SE	SP	AUC	PPP	NPP	K	SE	SP	AUC
1	F NWF	24	.551	.865	.761	.839	.545	.353	.538	.868	.787	.668	.791	.427	.660	.881	.289
	W NWF	50	.535	.747	.711	.766	.509	.260	.551	.831	.762	.618	.789	.393	.589	.000	.046
	S NWF	50	.314	.899	.709	.828	.458	.182	.347	.929	.783	.708	.742	.317	.246	.061	.003
	W ORF	20	.820	.690	.848	.804	.712	.513	.745	.809	.865	.659	.865	.537	.001	.000	.406
	S ORF	40	.743	.823	.859	.867	.674	.544	.688	.880	.876	.740	.851	.578	.036	.005	.390
2	F ORF	44	.847	.620	.848	.778	.721	.479	.864	.707	.892	.560	.923	.495	.465	.003	.036
	W ORF	68	.816	.653	.853	.787	.693	.474	.821	.775	.894	.612	.909	.543	.841	.000	.049
	S ORF	90	.837	.595	.839	.764	.699	.444	.882	.719	.893	.576	.934	.522	.059	.000	.010
3	F ORF	77	.906	.468	.846	.683	.794	.386	.829	.668	.827	.506	.904	.418	.001	.000	.484
	W ORF	92	.928	.468	.865	.688	.833	.410	.824	.642	.833	.486	.899	.386	.000	.000	.234
	S ORF	110	.906	.541	.859	.714	.817	.460	.824	.660	.834	.500	.901	.407	.000	.000	.352

Note. ROC = receiver operating characteristic; EL = students with limited English proficiency; BM = benchmark score; SE = sensitivity; SP = specificity; AUC = area under the curve; PPP = positive predictive power; NPP = negative predictive power; K = kappa; F = fall; W = winter; S = spring; NWF = Nonsense Word Fluency; ORF = Oral Reading Fluency.

Table 5
ROC Two-Proportions Test Results for the Disability Status Disaggregation Analysis

		SwD							non-SwD						*p*		
Grade	Measure	BM	SE	SP	AUC	PPP	NPP	K	SE	SP	AUC	PPP	NPP	K	SE	SP	AUC
1	F NWF	24	.723	.714	.755	.825	.581	.415	.520	.868	.779	.720	.734	.408	.000	.000	.575
	W NWF	50	.708	.686	.741	.807	.558	.374	.522	.822	.753	.655	.725	.355	.000	.000	.787
	S NWF	50	.523	.857	.760	.872	.492	.325	.308	.927	.753	.731	.672	.260	.000	.010	.873
	W ORF	20	.846	.571	.794	.786	.667	.432	.768	.795	.869	.710	.841	.556	.033	.000	.029
	S ORF	40	.831	.714	.841	.844	.694	.541	.696	.875	.875	.785	.815	.584	.000	.000	.317
2	F ORF	44	.947	.514	.859	.810	.818	.523	.822	.702	.867	.604	.891	.497	.000	.000	.810
	W ORF	68	.947	.571	.868	.819	.826	.550	.794	.765	.876	.647	.874	.533	.000	.000	.810
	S ORF	90	.933	.457	.841	.788	.762	.446	.845	.711	.870	.614	.898	.516	.000	.000	.384
3	F ORF	77	.968	.370	.822	.729	.870	.397	.831	.647	.827	.519	.891	.413	.000	.000	.881
	W ORF	92	.968	.352	.825	.723	.864	.377	.825	.622	.831	.505	.892	.393	.000	.000	.857
	S ORF	110	.946	.352	.804	.718	.792	.348	.822	.653	.835	.526	.894	.424	.000	.000	.342

Note. ROC = receiver operating characteristic; SwD = students with disabilities; BM = benchmark score; SE = sensitivity; SP = specificity; AUC = area under the curve; PPP = positive predictive power; NPP = negative predictive power; K = kappa; F = fall; W = winter; S = spring; NWF = Nonsense Word Fluency; ORF = Oral Reading Fluency.

Table 6
ROC Two-Proportions Test Results for the Race Ethnicity Disaggregation Analysis: Hispanic to White

			Hispanic						White						*p*		
Grade	Measure	BM	SE	SP	AUC	PPP	NPP	K	SE	SP	AUC	PPP	NPP	K	SE	SP	AUC
1	F NWF	24	.547	.831	.761	.795	.604	.366	.640	.869	.845	.640	.871	.511	.001	.069	.000
	W NWF	50	.574	.799	.739	.774	.609	.363	.622	.817	.808	.548	.857	.420	.097	.435	.005
	S NWF	50	.329	.896	.743	.790	.526	.212	.433	.930	.828	.689	.820	.412	.000	.038	.000
	W ORF	20	.829	.703	.862	.769	.773	.535	.823	.815	.897	.616	.928	.577	.787	.000	.066
	S ORF	40	.738	.843	.859	.849	.728	.573	.756	.878	.911	.693	.910	.618	.478	.084	.005
2	F ORF	44	.837	.659	.844	.723	.800	.506	.894	.697	.909	.509	.949	.475	.007	.180	.001
	W ORF	68	.801	.685	.847	.729	.767	.490	.856	.780	.907	.576	.939	.548	.018	.000	.002
	S ORF	90	.829	.655	.841	.719	.791	.493	.925	.712	.911	.529	.964	.512	.000	.042	.000
3	F ORF	77	.876	.594	.846	.635	.855	.453	.837	.662	.830	.443	.883	.302	.085	.030	.503
	W ORF	92	.900	.554	.859	.620	.873	.436	.837	.649	.838	.434	.925	.369	.004	.003	.368
	S ORF	110	.886	.602	.859	.643	.867	.471	.829	.673	.842	.449	.924	.389	.012	.023	.465

Note. ROC = receiver operating characteristic; BM = benchmark score; SE = sensitivity; SP = specificity; AUC = area under the curve; PPP = positive predictive power; NPP = negative predictive power; K = kappa; F = fall; W = winter; S = spring; NWF = Nonsense Word Fluency; ORF = Oral Reading Fluency.

7. Three different patterns of AUCs emerged: for White students, the AUCs met the criterion for all measurements; for Hispanic students, the AUCs met the criterion for all ORF measurements, but not NWF; for American Indian students, the AUCs met the criterion for ORF in Grades 1 and 2 only, but not for NWF. For both White and Hispanic students, SE met the criterion in Grades 2 and 3 as well as winter ORF in Grade 1. For the American Indian students, SE only met the criterion for spring ORF in Grade 2 and fall ORF in Grade 3. SP met the criterion for all three groups at all measurements in Grade 1 except for winter ORF for the Hispanic students. No measurements in Grades 2 or 3 met the criterion for SP. Because the comparisons can only be conducted between two groups, Hispanic and American Indian students were compared to White students separately.

In comparing Hispanic students to White students, significant differences in AUC were found for fall and spring NWF in Grade 1 and fall and spring ORF in Grade 2. A similar pattern was noted for SE with the exception of fall ORF in Grade 2. SP was significantly different between the groups for winter ORF in Grades 1 and 2. There were no significant comparisons in Grade 3.

In comparing American Indian students to White students, only fall NWF met the criterion for significance. For SE, all comparisons in Grade 1 met the criterion for significance, as did fall ORF in Grade 2. For SP, all three measurements in Grade 3 met the criterion for significance.

Summary of ROC results. Across the four disaggregation groups, there were few statistically significant differences in AUC between the groups being compared. All of these differences were with the NWF measure with the exception of ORF at grade 2 comparing Hispanic and White students. When examining the SE and SP indexes, there tended to be more differences in SE in Grade 1, particularly with the NWF measure; and more differences in SP in Grades 2 and 3 (most clearly illustrated in the American Indian to White comparisons). A final pattern was that for some groups, at certain grade levels (EL Grade 3, SwD Grades 2 and 3) there were significant differences in SE and SP, but not in AUC.

Quantile Regression Analyses

To answer the second research question, "How much does the accuracy of prediction of the NWF and ORF measures of the DIBELS on a state criterion-referenced test vary as a function of level of performance when examined across the disaggregation categories of NCLB?," a series of quantile regression models were developed and the graphs of the resultant correlations plotted and compared. Interpretation of the graphs was conducted by visual inspection comparing the regression plots between the groups. Floor or ceiling effects are demonstrated when a line is not horizontal (horizontal lines indicating the regression coefficients are similar across all points in the performance range). When two groups have similar plots, this indicates that the floor or ceiling effect (or lack thereof) affects both groups similarly. When the plots are different, one group is affected by the floor/ceiling effect more than the other.

Economic disadvantage. As can be seen in Figure 1, there are differences in the plots for the fall and winter NWF and ORF at Grades 1 and 2. The spring measurements in Grades 1 and 2 as well as all three measurements in Grade 3 have similar plots for both groups.

Limited English proficiency. As can be seen in Figure 2, there are differences in the plots for spring NWF and winter ORF in Grade 1 as well as fall and winter ORF in grade 2. All three measurements in Grade 3 have similar plots for both groups.

Disability status. As can be seen in Figure 3, there are differences in the plots for nearly every measurement point except spring NWF in Grade 1 and fall ORF in Grade 3.

Race/Ethnicity. As can be seen in Figure 4, there are differences among the plots for fall and winter NWF and winter ORF in Grade 1 as well as fall ORF in Grade 2. There

Table 7
ROC Two-Proportions Test Results for the Race Ethnicity Disaggregation Analysis: American Indian to White

		American Indian							White						*p*		
Grade	Measure	BM	SE	SP	AUC	PPP	NPP	K	SE	SP	AUC	PPP	NPP	K	SE	SP	AUC
1	F NWF	24	.283	.894	.723	.800	.432	.137	.640	.869	.845	.640	.871	.511	.000	.478	.001
	W NWF	50	.233	.915	.735	.813	.423	.115	.622	.817	.808	.548	.857	.420	.000	.017	.073
	S NWF	50	.133	.979	.715	.875	.407	.074	.433	.930	.828	.689	.820	.412	.000	.059	.004
	W ORF	20	.450	.894	.822	.897	.508	.317	.823	.815	.897	.616	.928	.577	.000	.055	.020
	S ORF	40	.567	.915	.837	.917	.569	.435	.756	.878	.911	.693	.910	.618	.001	.280	.016
2	F ORF	44	.714	.795	.862	.829	.774	.604	.894	.697	.909	.509	.949	.475	.000	.049	.124
	W ORF	68	.794	.773	.873	.829	.774	.604	.856	.780	.907	.576	.939	.548	.147	.873	.267
	S ORF	90	.825	.727	.859	.789	.786	.570	.925	.712	.911	.529	.964	.512	.013	.757	.087
3	F ORF	77	.855	.492	.765	.676	.696	.357	.837	.662	.830	.443	.883	.302	.617	.000	.093
	W ORF	92	.764	.475	.743	.686	.680	.360	.837	.649	.838	.434	.925	.369	.095	.000	.012
	S ORF	110	.745	.492	.733	.676	.654	.327	.829	.673	.842	.449	.924	.389	.063	.000	.004

Note. ROC = receiver operating characteristic; BM = benchmark score; SE = sensitivity; SP = specificity; AUC = area under the curve; PPP = positive predictive power; NPP = negative predictive power; K = kappa; F = fall; W = winter; S = spring; NWF = Nonsense Word Fluency; ORF = Oral Reading Fluency.

are also differences for the Hispanic group in spring ORF for Grade 2 (on the higher end of the distribution) and the American Indian group in fall ORF for Grade 3 (again, particularly on the high end). All other plots are similar.

Summary of quantile regression results. In general, there was less bias in predictive validity in Grade 3 than in Grades 1 and 2 as well as less of an influence of a floor effect (i.e., slope in the regression line). The patterns in Grade 1 appeared similar for NWF and ORF except for the EL comparisons. However, it should be noted that although there are many differences in performance of different groups across measures, there are also many similarities—indicating that patterns of potential bias are not extreme or consistent.

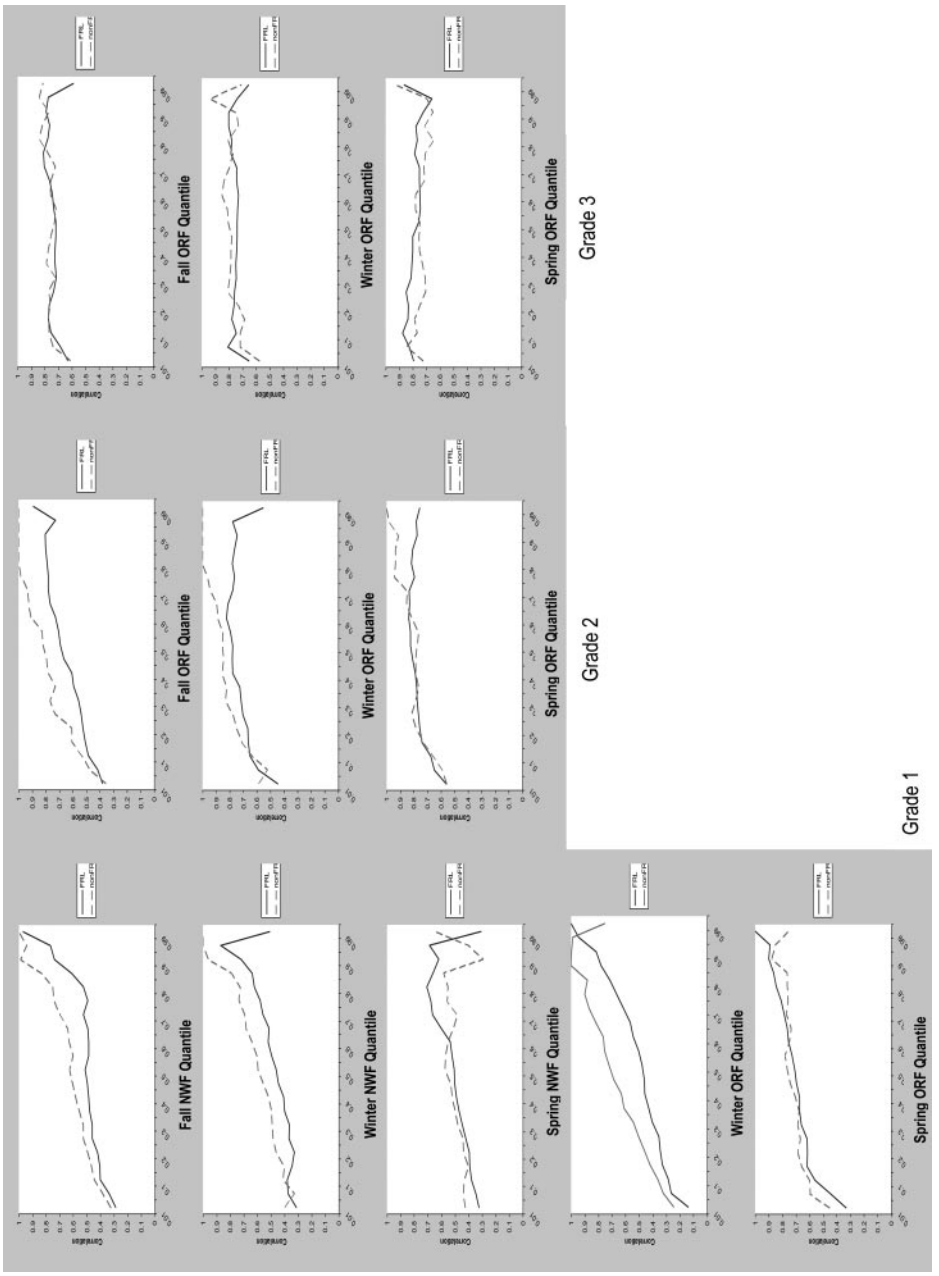

Figure 1. Quantile regression plots for grades 1–3 *DIBELS* measures for the economic disadvantage disaggregation analysis.

Discussion

In our nation's push to improve educational outcomes for all students, examination of bias in predictive validity of educational measures is vital. Consistency in our decision making is important in order to ensure consistency in service delivery and outcomes (Barnett et al., 2007) and to prevent over- or underidentification of a subgroup of students. Unfortunately, studies of bias in predictive validity of screening measures are relatively uncommon (Betts et al., 2008). In this study, we examined universal screening data for bias in predictive validity across the disaggregation categories mandated by NCLB. We found that measures with good overall predictive validity

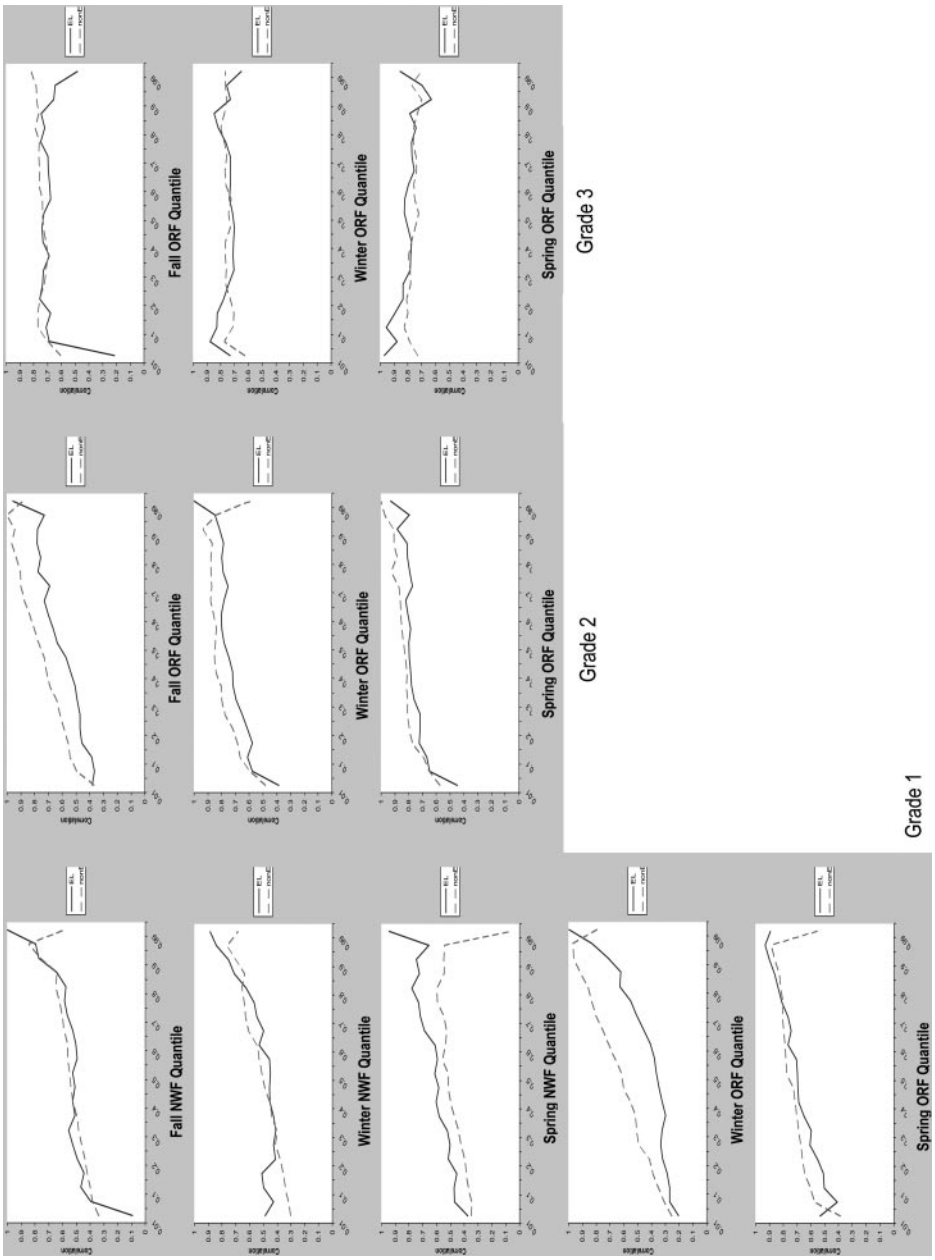

Figure 2. Quantile regression plots for Grade 1–3 DIBELS measures for the limited English proficiency disaggregation analysis.

(NWF, ORF) may not demonstrate consistent levels of predictive validity when focusing on different subgroups. Our results also suggest that this differential prediction varies across the subgroup analyses. Findings support prior research in which the patterns of predictive validity (or bias) have varied across studies.

There are many potential explanations for this variation in pattern across studies. In addition to the typical differences across research studies (different settings, participants, and so on), studies vary in use of criterion measures, inclusion and exclusion of variables, and instruction and intervention. Because studies of prediction bias involve relating a predictor measure to an outcome measure and use of a cut score, each of these components may contribute to differential pre-

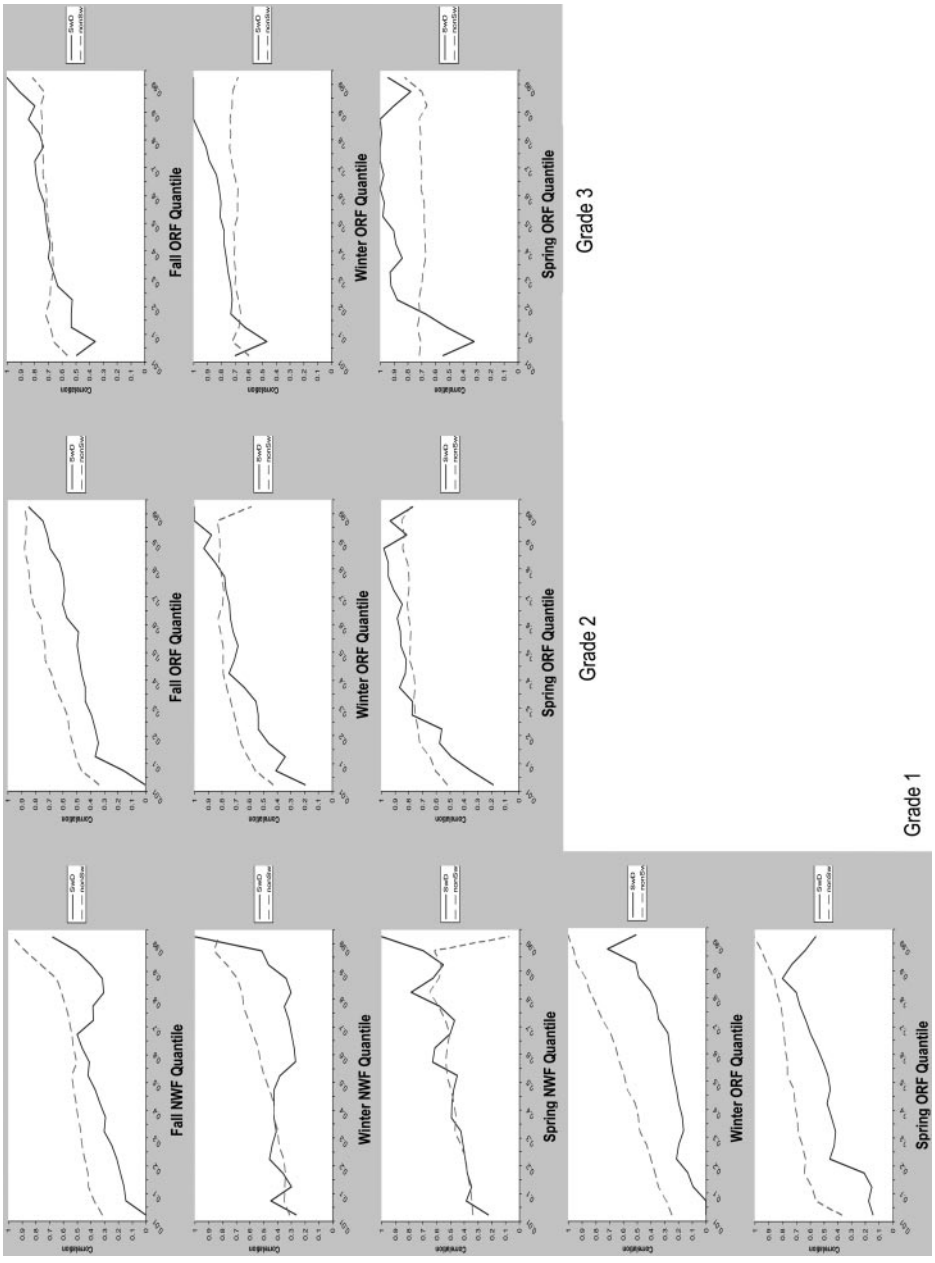

Figure 3. Quantile regression plots for Grade 1–3 DIBELS measures for the disability status disaggregation analysis.

diction patterns. The bias may be conceived as residing with the predictor measure (i.e., it being the dominant factor influencing the differential prediction because of variation in performance or functioning), the criterion measure, or the cut scores for one measure or the other (Flaugher, 1978). Results of predictive validity bias studies can indicate the presence of differential prediction, but generally not the location of that bias. A second source of variation is differential inclusion of variables. If a variable is correlated with both the predictor and the criterion measures, the coefficients (and therefore the decisions) may be biased because of the omission of the variable rather than the performance of the measures (Johnson, Carter, Davison, & Oliver, 2001). This could be the influencing factor involved in the

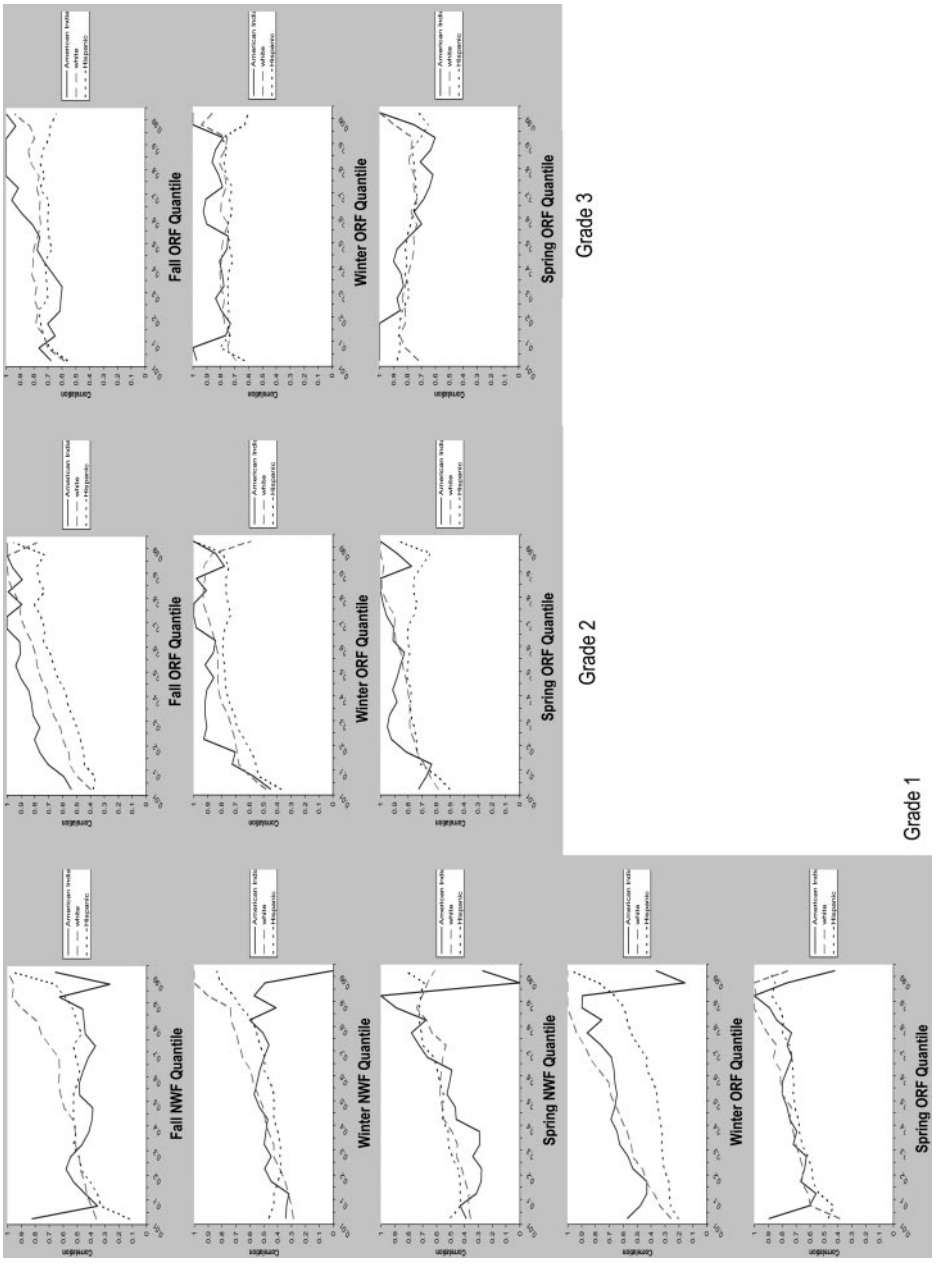

Figure 4. Quantile regression plots for Grade 1–3 DIBELS measures for the race/ethnicity disaggregation analysis.

participants in different studies receiving different instruction and intervention. Because the predictive validity studies involve a 3- to 6- month lag between administration of the predictor and criterion measures, participants received instruction and most likely differential intervention based on individual needs. This instruction may vary across studies and across classrooms or schools within studies adding another source of variance. These considerations highlight the importance of examining a phenomenon across studies to examine the pattern in greater detail.

In relation to the findings of Catts et al. (2009), we found a similar pattern of greater floor effects in Grade 1 (for both NWF and ORF) than in Grade 2, with little to no floor effect in Grade 3. As students progress in grade, their performance distribution is high enough to not have the restriction of range indicative of a floor effect. As far as differential floor effects across the disaggregation categories, there was not a clear pattern. All groups demonstrated floor effects at the fall and winter screenings for NWF and the winter ORF screening in Grade 1. All but the race/ethnicity comparisons also demonstrated floor effects in fall of Grade 2. No subgroup demonstrated floor effects in the spring of any year (i.e., potential concurrent predictive bias). No group demonstrated floor effects at all measurement points across all grades. The pattern displayed in this study appears to be that the first measurement period within a year (Grades 1 and 2) is the most likely to exhibit differential prediction. This makes sense in that group performance at the beginning of the year is likely to be lower than at other points, making the impact of floor effects more likely. If two groups perform differently, one would be more susceptible to floor effects than the other (i.e., the group with the lower overall performance). In addition, if the groups came into their education with different prior knowledge and experience, or were receiving differential curriculum or instruction (or differentially effective curriculum or instruction), different levels or patterns of performance could be expected (Donovan & Cross, 2002). The pattern of differential prediction in this study potentially caused by floor effects in the lower grades was not duplicated in the ROC analyses.

From the ROC analyses, although AUC is a valid effect size statistic for use in comparisons (Swets, 1988), the present results suggest that it may not be best to use it in isolation to judge bias in predictive validity. Differences in both SE and SP between two groups, with decisions for one group having higher SE and decisions for the other having higher SP, can actually offset each other in the determination of AUC. The clearest examples of this phenomenon are in Grade 3 for the limited English proficiency comparisons (Table 4) and all grades for the disability status comparison (Table 5). For these comparisons, AUC was similar between the groups, but there were significant differences between SE and SP. An implication of this pattern is that there are different mistakes in terms of decision making being made for different groups of students. From our results for the limited English proficiency comparison, ORF at Grade 3 demonstrated greater sensitivity for EL (i.e., the measures were better at identifying which individuals in the EL group would not meet proficiency on the outcome than for the non-EL group). Conversely, both measures demonstrated better specificity for non- EL (i.e., the measures were better at identifying which individuals in the non-EL group would meet or exceed the criterion for proficiency on the outcome measure than for the EL group). If using ORF at Grade 3 to screen students using a direct route approach, wherein those predicted to *not* meet the proficiency criterion on the outcome measure are automatically placed in supplemental instruction, or Tier 2 (Jenkins et al., 2007), one would make more false-positive errors for the EL group. This means that more EL students would be placed into Tier 2 intervention programs that they do not necessarily require than their non-EL peers. The reverse would also occur: more non-EL students would *not* receive Tier 2 services that they needed than their EL peers (i.e., a higher false-negative rate for non-EL).

One way to examine the presence of counteracting SE and SP is to use multiple

indexes of classification accuracy. If one index indicates a difference (e.g., SE), yet another does not (e.g., SP), there is a different pattern of predictive validity than if both indexes demonstrate differences. In the case of a nonsignificant AUC with significant SE and SP, this would also provide a check of whether there are two phenomena counteracting each other (indicating that the phenomenon may be a result of differential base rates on the criterion measure). Another strategy for examining and combating differential predictive validity is to identify different cut scores for different groups and/or different outcome measures (Roehrig et al., 2007). By systematically identifying different cut scores that maintain the same levels of sensitivity and specificity, similarity in proportions of false positives and false negatives across different groups could be ensured. However, one potential caution would be that generalizability of performance can be lost for the decisions made for individual students. If each school or district uses different criteria to determine the direct route to supplemental or intensive services, a student could move from one building where he was predicted to be proficient on the outcome measure (and therefore not receiving supplemental services) to another where he was not predicted to be proficient and therefore in need of additional instruction. Although both decisions may be correct in determining the student's needs, it provides an additional level of coordination and judgment for the school to which the student moves.

In addition to the need to examine multiple indexes in a determination of differential predictive validity, there are other implications from our results. First is that individual students are included in multiple groups (e.g., a Latino student receiving free lunch, or a student with a disability with limited English). As such, if different cut scores are developed for use with different groups, which one would be used for making a decision about service delivery for an individual student? Because the disaggregation categories are mutually exclusive (e.g., a student cannot be both economically disadvantaged and not economically disadvantaged) and comprehensive (e.g., all students are either economically disadvantaged or not), every student can be classified along the dimensions of every disaggregation category. This would mean up to four separate cut scores (five if you add sex) for every student as identifying which one was the most accurate would be a difficult web to untangle.

A second implication is that the absence of prediction bias does not automatically equal fairness. As stated previously, predictive validity has to do with prediction of outcomes; that bias in predictive validity occurs when a test differentially predicts that outcome for one group as compared to another. By contrast, fairness can be conceived as differences in the mean test scores that an individual is being compared to (or used to develop the criterion) that are not directly related to the focus of the measure (i.e., construct-irrelevant variance). Although approaches to quantitatively address lack of fairness in assessment do exist (see Helms, 2006), they are not widely adopted or used.

Limitations

The above findings should be interpreted in light of some potential limitations. First, the sizes of groups being compared were not equivalent, which could lead to differences in the consistency of scores and error for the groups (Tabachnick & Fidell, 2007). Second, some of the group sizes were relatively small ($n > 100$ students). Although each group included was large enough to run the analyses, larger samples would provide more stable estimates—particularly for the quantile regression analyses in which more stable estimates would provide smoother plot lines (Koenker, 2005). A third limitation is the lack of an African American subgroup in the race/ethnicity analyses. This is a potential limitation because African American students are, and have traditionally been, one of the larger racial/ethnic groups in the United States. Fourth, the data for this study came exclusively from Reading First schools. As such, the minority, English learner, and economically disadvantaged proportions of students are higher than those of the state as a whole. Similar to the

results from the Catts et al. (2009) study, the extent of this effect in unclear. Last, as with most any study examining prediction with screening measures, there is an intervening agent present as the results from the screening measures were intended to be used to make decisions about placement and intervention provision. This can affect the classification accuracy estimates despite the fact that it is precisely the purpose for which the measures are designed (Hosp, Dole, & Hosp, 2006).

Implications for the Practice of School Psychology

Despite the above-mentioned limitations, there are messages that school psychologists can take from the current findings. First, use of a single measure is not prudent for screening decisions. Given these preliminary findings of bias in predictive validity, using other pieces of data to validate the decision from a screening measure should reduce the potential for false positives or negatives. Triangulation of data either from other screening measures that address the same skill in a different way or inclusion of progress monitoring data after the screening provides additional pieces of information with which to make a decision. Second, using a team to make decisions can be useful for screening decisions as well as eligibility decisions as mandated by Individuals with Disabilities Education Act (2004). Similar to use of multiple pieces of data or a structured decision-making process, it introduces an added layer of accountability to make sure that there is agreement in the decisions.

References

Barnett, D. W., Hawkins, R., Prasse, D., Graden, J., Nantais, M., & Pan, W. (2007). Decision-making validity in response to intervention. In S. R. Jimerson, M. Burns, & A. VanDerHeyden (Eds.), *Handbook of response to intervention: The science and practice of assessment and intervention* (pp. 106–116). New York: Springer.

Batsche, G., Elliott, J., Graden, J. L., Grimes, J., Kovaleski, J. F., Prasse, D., et al. (2005). *Response to intervention: Policy considerations and implementation.* Alexandria, VA: National Association of State Directors of Special Education.

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.

Buck, J., & Torgesen, J. (2002). *The relationship between performance on a measure of oral reading fluency and performance on the Florida Comprehensive Assessment Test* (FCRR Technical Report No. 1). Tallahassee: Florida Center for Reading Research.

Carran, D. T., & Scott, K. G. (1992). Risk assessment in preschool children: Research implications for the early detection of educational handicaps. *Topics in Early Childhood Special Education, 12,* 196–211.

Catts, H. W., Fey, M. E., Zhang, X., & Tomblin, J. B. (2001). Estimating the risk of future reading difficulties in kindergarten children: A research-based model and its clinical implementation. *Language, Speech, and Hearing Services in Schools, 32,* 38–50.

Catts, H. W., Petscher, Y., Schatschneider, C., Bridges, M. S., & Mendoza, K. (2009). Floor effects associated with universal screening and their impact on the early identification of reading difficulties. *Journal of Learning Disabilities, 42,* 162–176.

Cleary, T., Humphreys, L. G., Kendrick, S. A., & Wesman, A. (1975). Educational uses of tests with disadvantaged students. *American Psychologist, 30,* 15–41.

Cole, N., & Moss, P. (1993). Bias in test use. In R. L. Linn (Ed.), *Educational measurement* (3rd ed.; pp. 201–220). Phoenix, AZ: The Oryx Press.

Donovan, M. S., & Cross, C. T. (Eds.). (2002). *Minority students in special and gifted education.* Washington DC: National Academy Press.

Drew, C. J., Hardman, M. L., & Hosp, J. L. (2008). *Designing and conducting research in education.* New York: Sage.

Fien, H., Baker, S. K., Smolkowski, K., Mercier-Smith, J. L., Kame'enui, E. J., & Beck, C. T. (2008). Using nonsense word fluency to predict reading proficiency in kindergarten through second grade for English learners and native English speakers. *School Psychology Review, 37,* 391–408.

Flaugher, R. L. (1978). The many definitions of test bias. *American Psychologist, 33,* 671–679.

Foorman, B. F., Francis, D. J., Fletcher, J. M., Schatschneider, C., & Mehta, P. (1998). The role of instruction in learning to read: Preventing reading failure in at-risk children. *Journal of Educational Psychology, 90,* 37–55.

Good, R. H., Kaminski, R. A., Shinn, M., Bratten, J., Shinn, M., Laimon, D., et al. (2004). *Technical adequacy of DIBELS: Results of the Early Childhood Research Institute on measuring growth and development* (Technical Report No. 7). Eugene: University of Oregon.

Good, R. H., Kaminski, R. A., Smith, M. R., & Bratten, J. (2001). *Technical adequacy and second grade DIBELS Oral Reading Fluency (DORF) passages* (Technical Report No. 8). Eugene: University of Oregon.

Haladyna, T. M. (2006). Roles and importance of validity studies in test development. In S. M. Downing & T. M. Haladyna (eds.), *Handbook of test development* (pp. 739–758). Hillsdale, NJ: Lawrence Erlbaum Associates.

Harn, B. A., Stoolmiller, M., & Chard, D. J. (2008). Measuring the dimensions of alphabetic principle on the reading development of first graders: The role of

automaticity and unitization. *Journal of Learning Disabilities, 41,* 143–157.
Helms, J. E. (2006). Fairness is not validity or cultural bias in racial-group assessment: A quantitative perspective. *American Psychologist, 61*(8), 859–870.
Hintze, J. M., & Silberglitt, B. (2005). A longitudinal examination of the diagnostic accuracy and predictive validity of R-CBM and high-stakes testing. *School Psychology Review, 34,* 372–386.
Hoover, H. D., Dunbar, D. A., Frisbie, D. A., Oberley, K. R., Bray, G. B., Naylor, R. J. (2003). *The Iowa Tests of Basic Skills.* Rolling Meadows, IL: The Riverside Publishing Company.
Hosp, J. L., & Ardoin, S. (2008). Assessment for instructional planning. *Assessment for Effective Intervention, 33,* 69–77.
Hosp, J. L., Dole, J. A., & Hosp, M. K. (2006, July). *DIBELS as a predictor of proficiency on high stakes outcome assessments for at-risk readers.* Paper presented at the annual meeting of the Society for the Scientific Study of Reading, Vancouver, BC.
Hosp, J. L., & Reschly, D. J. (2003). Referral rates for intervention or assessment: A meta-analysis of racial differences. *The Journal of Special Education, 37,* 67–80.
Hughes, C., & Dexter, D. (2007). *Universal screening within a response-to-intervention model* (report brief for RTI Action Network). New York: National Center for Learning Disabilities.
Ikeda, M. J., Neessen, E., & Witt, J. C. (2008). Best practices in universal screening. In A. Thomas & J. Grimes (Eds.), *Best practices in school psychology* (5th ed., Vol. 2, pp. 103–114). Bethesda, MD: National Association of School Psychologists.
Individuals with Disabilities Education Improvement Act, 20 U.S.C. § 1400 (2004).
Jenkins, J. R., Hudson, R. F., & Johnson, E. S. (2007). Screening for at-risk readers in a response to intervention framework. *School Psychology Review, 36,* 582–600.
Johnson, J. W., Carter, G. W., Davison, H. K., & Oliver, D. H. (2001). A synthetic validity approach to testing differential prediction hypotheses. *Journal of Applied Psychology, 86,* 774–780.
Koenker, R. (2005). *Quantile regression.* New York: Cambridge University Press.
Metz, C. E. (1978). Basic principles of ROC analysis. *Seminars in Nuclear Medicine, 8,* 283–298.
No Child Left Behind Act, 20 U.S.C. § 6301 (2002).
O'Connor, R. E., & Jenkins, J. R. (1999). Prediction of reading disabilities in kindergarten and first grade. *Scientific Studies of Reading, 3,* 159–197.
Race to the Top, 26 U.S.C. § 1 (2009).
Rampsey, B. D., Dion, G. S., & Donahue, P. L. (2009). *NAEP 2008 trends in academic progress* (NCES 2009–479). Washington, DC: National Center for Education Statistics, Institute for Education Sciences, U.S. Department of Education.
Renaissance Learning. (2011). STAR Reading. Retrieved from http://www.renlearn.com/sr/
Ritchey, K. D. (2008). Assessing letter sound knowledge: A comparison of letter sound fluency and nonsense word fluency. *Exceptional Children, 74,* 487–506.
Ritchey, K. D., & Speece, D. L. (2004). Early identification of reading disabilities: Current status and new directions. *Assessment for Effective Intervention, 29*(4), 13–24.
Roehrig, A. D., Petscher, Y., Nettles, S. M., Hudson, R. F., & Torgesen, J. K. (2007). Accuracy of the DIBELS Oral Reading Fluency measure for predicting third grade reading comprehension outcomes. *Journal of School Psychology, 46,* 343–366.
Salvia, J., Ysseldyke, J. E., & Bolt, S. (2009). *Assessment: In special and inclusive education* (11th ed.). New York: Wadsworth.
Shanahan, T. (2003). Review of the DIBELS: Dynamic Indicators of Basic Early Literacy Skills (6th ed.). In B. S. Plake, J. C. Impara, & R. A. Spires (eds.), *The sixteenth mental measurements yearbook* (pp. 310–313). Lincoln, NE: Buros Institute of Mental Measurements.
Sprinthall, R. C. (2003). *Basic statistical analysis* (7th ed.). New York: Pearson.
Stage, S. A., & Jacobson, M. D. (2001). Predicting student success on a state-mandated performance-based assessment using oral reading fluency. *School Psychology Review, 30,* 407–419.
Swets, J. A. (1988). Measuring the diagnostic accuracy of diagnostic systems. *Science, 240,* 1285–1293.
Swets, J. A. (1996). *Signal detection theory and ROC analysis in psychology and diagnostics.* Hillsdale, NJ: Lawrence Erlbaum Associates.
Tabachnick, B. G., & Fidell, L. S. (2007). *Using multivariate statistics* (5th ed.). Needham Heights, MA: Allyn & Bacon.
Utah State Office of Education. (2007). *Utah ELA CRT technical manual.* Salt Lake City, UT: Author. Available at www.usoe.k12.ut.us
Wiley, H. I., & Deno, S. L. (2005). Oral reading and maze measures as predictors of success for English learners on a state standards assessment. *Remedial and Special Education, 26,* 207–214.
Woodcock, R. (1998). *Woodcock Reading Mastery Test—Revised/Normative Update.* Circle Pines, MN: American Guidance Service.

Date Received: November 11, 2010
Date Accepted: January 15, 2011
Action Editor: Sandy Chafouleas ■

John L. Hosp, PhD, is an associate professor of teaching and learning at the University of Iowa and codirector of the Center for Disability Research and Education. His current research interests include aligning assessment and instruction through curriculum-based measurement and curriculum-based evaluation, particularly in the elementary grades, as well as the disproportionate representation of students of color in special education.

Michelle K. Hosp, PhD, is a consultant with the Iowa Department of Education. Her interests are curriculum-based measurement and curriculum-based evaluation for reading and literacy with elementary students. She has extensive experience writing about reading and assessments as well as presenting at local, state, and national conferences. She is also currently a trainer for the National Center for Response to Intervention.

Janice A. Dole, PhD, is a professor of education at the University of Utah, where she teaches graduate courses in reading. Her research interests include school reform in reading, professional development, and summer reading loss in high-poverty schools.