2012-2013 DIBELS Data System Update
Part II: DIBELS Next Benchmark Goals

Summary

The DIBELS Data System (DDS) recommends the use of new benchmark goals for all recommended DIBELS Next measures beginning in the 2012-2013 school year. These new, recommended goals from the University of Oregon (UO) are different from the former goals from Dynamic Measurement Group, Inc. (DMG), and by and large they are considerably more ambitious. We understand that recommending different goals may create challenges for districts and schools; however, the UO recommended goals align more closely with other recent increases in state and federal standards requiring stronger literacy skills that generalize to higher-level comprehension processes.

Our findings are based on evidence from the following four key sources: (a) an external technical review of DIBELS Next materials, (b) an analysis of the procedures used to establish the DMG former goals, (c) consistent feedback from users, and (d) best practices in education research on sample selection and study replication. With these sources of information we found that, for two thirds of the individual DIBELS Next assessments, students who just meet the DMG former benchmark goals have less than a 60% chance of scoring at or above the 40th percentile on the Sat10 Total Reading Composite. In other words, many of these students will have failed to reach a healthy reading outcome, even when the performance on screening measures indicated otherwise. Given the implications of these findings, we concluded it was essential to move forward with two key actions:

(i) Provide information to all DDS customers on the accuracy of the DMG former goals based on our analysis, and
(ii) Provide a *choice* to all DDS customers by making available a second set of goals (i.e., UO recommended goals) to use in their schools.

The goals we recommend give priority to the *identification of students who are likely to be at risk for reading difficulties* while simultaneously using a stringent criterion (i.e., SAT10 performance at or above the 40th percentile) as the standard for benchmark reading performance. We base our predictions about students' overall reading performance on a large and nationally representative sample of students.

We want to emphasize that you have a choice in the benchmark goals you select with DIBELS Next. You can continue to use the DMG former goals if you wish, which the DDS will fully support, or you can use the UO recommended goals, which we also fully support. Because the goals you use for screening can so dramatically affect instructional planning, however, we do believe that using the UO recommended goals will allow you to more accurately estimate your students' current and future reading performance, and make the best decisions you can about providing support for students based on their need.

Introduction and Rationale

DIBELS Next was *first* made available to schools and supported on the DDS. The DDS was the first data system to support DIBELS Next, and we made the decision to do so because we felt it was important for DDS schools to be able to use the new measures if they so wished. Since 2010, our research team at the University of Oregon, Center on Teaching and Learning (UO CTL), which is a separate, independent, not-for-profit entity that is not affiliated with DMG, has been collecting feedback from schools about their use of DIBELS. At the same time, our researchers have been conducting a series of rigorous, national evaluations addressing the technical features of DIBELS Next. The results of this work, along with feedback from DDS users, led us to make some important changes to the DDS and the services we offer our customers. This Technical Brief is the second in a series of reports that will explain the rationale and research basis for the system advances we have instituted. For more information about changes to the DIBELS Next Composite Score, please see our first technical brief in this series, Part I: DIBELS Next Composite Score (UO CTL, 2012a).

Benchmark Goals as Criterion Levels of Performance. The process of establishing benchmark goals for DIBELS Next or any screening measure is critically important. Benchmark goals represent *criterion levels of performance;* these are performance standards based on an indicator measure that we believe reasonably predict successful, overall comprehensive reading skills. The *sample of students* used to determine these criterion levels of performance (e.g., benchmark, strategic, or intensive) is one major decision that must be very carefully considered in the development of any educational test. How these decisions are made can have a profound influence on how we determine whether students are reading successfully or are at risk for serious reading difficulty. Different procedures for sample selection and statistical analysis can result in dramatically different goals that differentially categorize the overall reading health *of the same student*. Thus, it is critical that we use the best tools that are available to establish benchmark goals—based on what we know about rigorous scientific research—and in this case, it is critical to include a representative sample of students and appropriate statistical procedures.

Rationale for New Goals. We established new recommended goals for four primary reasons.

First, we conducted an external technical review of DIBELS Next using the review criteria published by the National Center on Response to Intervention (NCRTI, 2011b). Using these criteria, the DMG former goals for *individual DIBELS Next measures* did *not* meet the validity standards specified by NCRTI because they were linked only to the DIBELS Next Composite Score—not to an *external* measure of overall reading performance.

Second, we analyzed the former DMG individual goals using a sample of DDS schools. When we examined the school and student sample on which the former DIBELS Next goals were based, it was clear that the sample was substantially different from the race, ethnicity, and socio-economic status of the population of students in the DDS *and* the U.S. population as a whole (see the Technical Brief Part I (UO CTL, 2012a), as well as p. 4 of this report for more details).

Third, we heard from schools, districts, and states that the DMG former goals were simply "inadequate." At the same time that we began to review the DIBELS Next Technical Manual (DMG, 2011), schools began reporting to us that many of their students who achieved benchmarks on DIBELS Next were not predicted to meet standards on their state tests. Users raised concerns that the DIBELS Next measures were not providing information of the same value as was provided by DIBELS 6th. Our analyses confirm these concerns, and indicate that the DMG former goals are likely to *misclassify* many students who actually need additional instructional support. The goals also *vary by grade level* in problematic ways. For example, the former DIBELS Next Oral Reading Fluency (ORF) goals for sixth graders are much lower than those for fifth graders—and this variation does not seem to be entirely explained by differences in passage difficulty (see DMG, 2012 for details).

Fourth, we agree with the statement in the DIBELS Next Technical manual (DMG, 2011) that "no single study can provide all the information necessary to evaluate generalizability (of benchmark goals). As with DIBELS 6th edition, multiple studies will evaluate the reliability, validity, and utility of DIBELS Next." (p. 48). Conducting high-quality research is a hallmark of the University of Oregon College of Education, and our research on the recommended benchmark goals for DIBELS Next is consistent with this standard.

In the sections that follow, we document the procedures we used to revise the DIBELS Next goals and cut points. For the sake of efficiency in this report, we include examples for two measures (i.e., Nonsense Word Fluency-Correct Letter Sounds (NWF-CLS) and ORF) and three grade-by-time point combinations (i.e., middle of kindergarten, beginning of Grade 3, and beginning of Grade 6). For the complete list of all recommended DIBELS Next goals, please see the DIBELS Next Recommended Goals Document (UO CTL, 2012b).

Method

Participants. The first important decision in the goal setting process is the size and composition of the sample of students who will be used in subsequent comparisons to other students in other schools and at other time points (*Standards for Educational and Psychological Testing,* AERA, APA, NCME, 1999). The sample of students used in the development of the UO recommended goals for the DIBELS Next measures was drawn from 28 schools in 15 states. Using a multistage, cluster sampling approach (O'Connell & McCoach, 2008), these 28 schools were selected from a larger sample of DDS schools. The number of students tested per grade ranged from 308 in grade 6 to 671 in Grade 1 (UO CTL, 2012a).

Equally, if not more important than sample size is the representativeness of the student group on which benchmark goals are based. The percent of students who qualify for free or reduced-price lunch in the sample used to calculate the UO recommended benchmark goals ranges from 59 to 64% across Grades K - 6. Similarly, the U.S. average percent of students who qualify for free or reduced-price lunch is 52% in each of Grades K - 6 (NCES, 2011). In contrast, the former goals as reported by DMG (2011) indicate an overall poverty rate (as indexed by the percent of students who qualify for free or reduced price lunch) of 16%.

The two samples also differ in terms of the race and ethnicity of the participants. Unlike the sample used by DMG for the former goals, the UO recommended goals sample *closely matches current U.S. demographic data* in terms of race and ethnicity (see Figure 1).

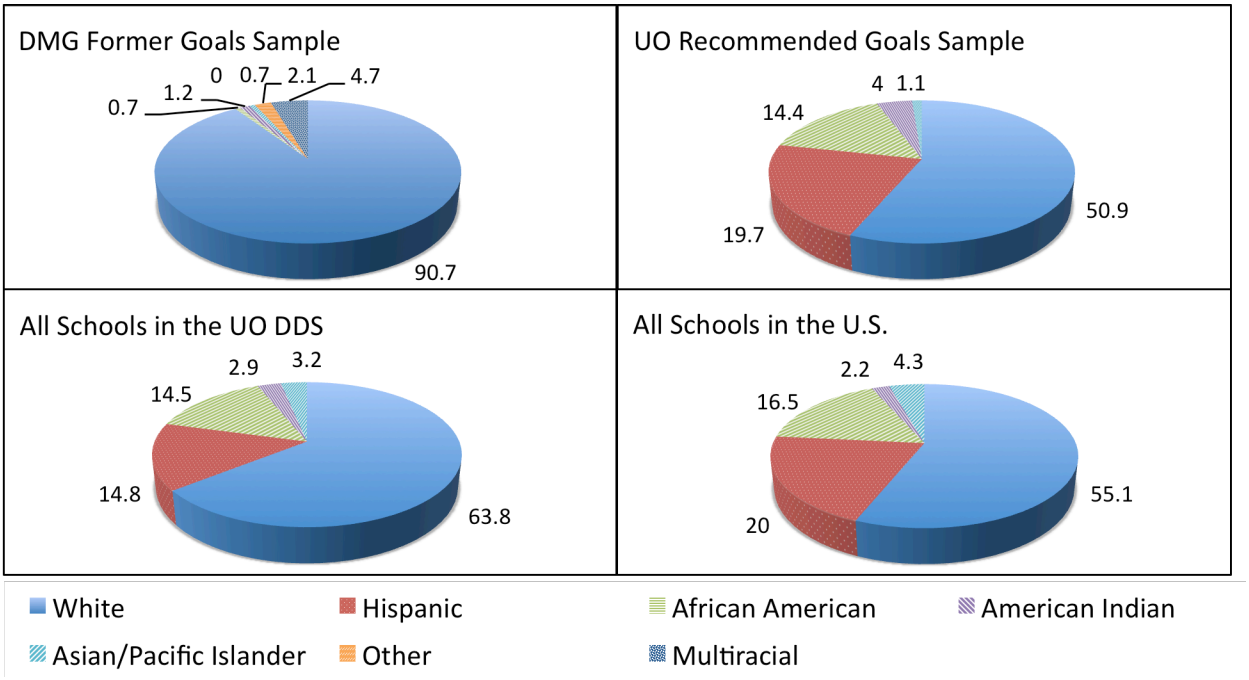

Figure 1. Comparison of participant samples used to develop DMG former and UO recommended benchmark goals relative to demographic data from all schools in the UO DDS and all public schools in the U.S. serving Grades K – 6. Note that, in the DMG former goals sample only, participants reported ethnicity (i.e., Hispanic or Not Hispanic) as a separate category from race. Thus, in addition to the racial categories listed in the upper left quadrant of this figure, 6% of the DMG former goals sample was also identified as having Hispanic or Latino/a ethnicity (see DMG, 2011).

The median proportion of students in our sample who are White is 50.9% (range across grades = 48.5 – 58.8%); the U.S. median is 55.1% (range = 54.6 - 55.9%). The median proportion of students in our sample who are Black is 14.4% (range = 12.2 – 17.7%); the U.S. median is 16.5% (range = 15.6 - 17.2). The median proportion of students in our sample who are Hispanic is 19.7% (range = 11.3 – 21.6%); the U.S. median is 20.0% (range = 18.9 - 20.4%). The median proportion of students in our sample who are American Indian/Alaska Native is 4% (range = 3.3 - 7.8%); the U.S. median is 2.2% (range = 2.1 – 2.8%). In contrast, the sample for the DMG former goals is identified as 91% White, with 1% or fewer students in all other categories except Multiracial (5%) and Hispanic or Latino/a (6%; DMG, 2011).

Criterion measure. Choosing an appropriate criterion measure is another important decision in the goal setting process. In establishing benchmark goals, the *Standards for Educational and Psychological Testing* (AERA, APA, NCME, 1999) require a link between student performance on screening measures (in our case, the individual DIBELS Next measures) and a standardized, widely used, *external criterion measure* of reading that represents an agreed-upon standard for acceptable student reading achievement. High-quality criterion measures in education should be developed and normed every seven years (Salvia & Ysseldyke, 2007), and include a

representative sample of students (AERA, APA, NCME, 1999). As an additional requirement, the NCRTI recommends that benchmark goals be linked to an outcome that occurs later in time (i.e., at least three months after screening; NCRTI, 2011b). As the external criterion measure for all grades, we chose the *Stanford Achievement Test—10th edition (SAT10; Pearson Education, Inc., 2004, 2007 Normative Update)* Total Reading Composite score, administered at the end of the school year.

In the DMG former goals study, the criterion measure in *some* analyses was the Group Reading Assessment and Diagnostic Evaluation (GRADE; Williams, 2001) test. Although the norm sample for the GRADE is outdated (having most recently been evaluated in 1998), the biggest concern is that, in the DMG former goals study, *individual DIBELS Next performance was never linked directly to the GRADE*. Rather, the goals for individual DIBELS Next measures were established via linking to the DIBELS Next Composite Score (DMG, 2011). This practice increases the error of prediction between any individual DIBELS Next goal and the ultimate outcome of interest: successful reading comprehension at the end of a school year on an external criterion measure of reading performance.

Statistical procedures. Conclusions drawn from the goal setting process can also differ substantially depending on the analytic approach used. We utilized two primary statistics in establishing the recommended DIBELS Next goals and cut points: (a) the Area Under the Curve (*AUC*) and (b) *sensitivity*.

The *AUC* is a measure of how accurately a test separates students being tested into the correct classifications or groups (Marzban, 2004; Peng & So, 2002; Pepe, 2003; Swets, 1996). If a test predicts perfectly how well a student will read at the end of the school year, the AUC will be equal to 1.0. If a test provides no information whatsoever about how well a student will read at the end of the school year (i.e., a coin flip determining good versus poor reading performance would do just as well), the AUC will be equal to .50.

The *sensitivity* indicator provides information about how well a given cut score on a given measure identifies students who have *not* met a criterion goal. It is expressed as a proportion. The sensitivity value represents the proportion of "truly" at-risk students who are correctly identified by the screener as being at risk. *Specificity* is the counterpart to sensitivity. Specificity is also expressed as a proportion, and represents the proportion of "truly healthy" readers who are accurately *not* identified as at risk by the screener (i.e., identified by the screener as "okay"). *Sensitivity can also be interpreted as the probability (likelihood) that a student who meets the criterion goal has been identified as such by the screener.*

In the analysis of the DMG former goals, we note that the primary statistic used to establish cut points is the negative predictive value (NPV). This statistic describes the probability of passing the criterion test among those who also pass according to the screener. At issue with NPV is that it is markedly affected by the prevalence of reading difficulty in the sample (Swets, 1996). Just as with sensitivity and specificity, the NPV associated with a perfect screener will be 1.0. However, unlike the values for sensitivity/specificity, *the NPV associated with a screener that carries no useful information is equal to 1 – the base rate* (i.e., prevalence of reading disability).

Consider a sample of students with a low prevalence of reading disabilities at .20. In this type of sample, the NPV for a useless screener would still look rather high at .80 (because NPV for a screener with no value in this sample = 1 – the base rate of .20). When the same cut point is evaluated with another sample that has a higher base rate (i.e., risk level for reading disability), it would be associated with a much lower NPV—assuming all else is equal. Thus, although the accuracy of the screener, defined by sensitivity and specificity, remains the same, the NPV can vary wildly even for a screener that provides no additional information. Incidentally, the Positive Predictive Value (PPV) has a similar flaw when used to choose an optimal cut point.

By utilizing a common outcome measure in all comparisons (i.e., the SAT10), as well as statistics that are not biased due to different prevalence rates of reading disability (Pepe, 2003), we believe we are able to offer much more stable and consistent goals for DIBELS Next across all grade levels and school populations.

Decision rules. For each measure at each time point that it is recommended, we calculated (a) the *benchmark goal*, and (b) the *cut point for risk*. Students with scores that are *at or above the benchmark goal* are labeled in green on DDS reports, with a descriptor that reads "likely to need core (instructional) support." We describe the benchmark goal in this way because it is the DIBELS score that most closely predicts scoring at or above the 40th percentile on the SAT10. The 40th percentile is widely used as an index above which it could be reasonably concluded a student is achieving in the average range or above. The benchmark goal is the lowest score on a DIBELS measure for which we predict with strong confidence that students will reach the criterion on the outcome measure.

We also established a "cut point for risk." Students with scores that are *below the cut point for risk* are labeled on DDS reports in red, with a descriptor that reads "likely to need intensive (instructional) support." We describe the cut point for risk in this way because it is the DIBELS score which, if a student is *below*, predicts a SAT10 score that is below the 20th percentile.

The *range* of scores between the benchmark goal and the cut point represents a group of scores for which we cannot offer a strong prediction of future performance. Students with scores that *are at or above the cut point for risk and below the benchmark goal* are labeled on DDS reports in yellow, with a descriptor that reads "likely to need strategic (instructional) support." Without intervention, students with scores in the strategic range are predicted to score between the 20th and the 39th percentile on the SAT10 at the end of the school year. However, many students with scores in the strategic range are also likely to score in the average range or above on the criterion measure. Because of this uncertainty, we recommend that schools strategically use resources to serve these students.

Analytic approach. The National Center for Response to Intervention (NCRTI, 2011b) recommends that criterion goals be linked to a single, well-established outcome. The Code of Fair Testing Practices in Education (JCTP, 2004) recommends that test developers should provide information to support recommended interpretations of results, including technical information and details about the nature of the comparison groups. In light of these and other test practices, *we chose to link each DIBELS Next measure at each time point (i.e., fall, winter and*

spring) to the SAT10 outcome measure: that is, the total reading composite percentile rank score in the spring. For example, in the fall of Grade 2, the raw scores of NWF-CLS, NWF Whole Words Read (NWF-WWR), ORF, Accuracy with ORF, and Retell were all linked to SAT10 end-of-year percentile rank. The winter and spring DIBELS scores from Grade 2 were also linked to the SAT10 end-of-year percentile rank.

We used a two-stage process for determining benchmark goals. First, we examined the validity of each measure using the AUC. Prior to conducting our analyses, we decided to only calculate benchmark goals for those measures with an AUC greater than .75. An AUC of less than .75 indicates that a measure does not represent accuracy beyond teacher judgment. We believe that providing benchmark goals for measures with AUC values that are less than .75 would imply greater confidence in the measures than is warranted.

Second, we conducted a diagnostic analysis of each measure. For each time point, we performed two sets of analyses: one to set the benchmark goal, and one to set the cut point for risk. For each analysis, we examined two statistics: sensitivity and specificity. We chose to focus on sensitivity and specificity (rather than PPV and NPV) because they *remain stable indicators regardless of the prevalence of reading difficulties in the population* (Pepe, 2003).

Further, we emphasized sensitivity in our analyses because of its practical application in a prevention model in education. Specifically, we want to be confident that students receive the instructional support they require as early as possible.

For benchmark goals, we identified the DIBELS Next score that corresponded to a sensitivity criterion of .90 (i.e., 90%). This criterion roughly corresponds to the statement that, *we will miss an opportunity to provide additional, strategic support to only 10% of students who are likely to score below the 40th percentile on the SAT10*.

We used a slightly different criterion when selecting cut points for risk. Our rationale for using a different criterion for the lower cut point has to do with the consequences associated with each type of decision. We are willing to "miss" more students with the cut points for risk because students who are not selected for intensive intervention (i.e., by not scoring at or below the cut point for risk) are still targeted for strategic intervention (i.e., by scoring below the benchmark goal). Thus, for the cut point for risk, we identified the DIBELS Next score that corresponded to a sensitivity criterion of .80 (i.e., 80%). This criterion roughly corresponds to the statement that, *we will miss an opportunity to provide intensive support to 20% of students who are likely to score below the 20th percentile on the SAT10*.

Results & Discussion

Value of individual DIBELS Next measures for universal screening. In Table 1 we provide a summary of the AUC for all DIBELS Next measures at appropriate benchmark testing occasions for Grades K through 6. Values in parentheses indicate an AUC of less than .75 for a particular measure at a particular timepoint in a particular grade (poor). Values in plain text indicate an AUC between .75 and .84 (good), values in plain text and shaded indicate AUC between .85 and .90 (very good), and values in bold text and shaded indicate AUC above .90 (excellent).

Table 1

AUC For DIBELS Next Measures Predicting Spring SAT10 Benchmark Performance (≥ 40th)

	AUC				AUC		
	Beg	Middle	End		Beg	Middle	End
Grade K				**Grade 3**			
FSF	.76	.75	–	ORF-WRC	.84	.85	.85
LNF	.80	.79	.78	ORF-Acc	.80	.80	.80
PSF	–	.75	(.68)	Retell	.75	(.69)	(.73)
NWF-CLS	–	.80	.80	Retell-Q	na	na	na
NWF-WWR	–	(.68)	.79	Daze	.80	.81	.80
Grade 1				**Grade 4**			
LNF	.77	–	–	ORF-WRC	.82	.83	.81
PSF	(.60)	–	–	ORF-Acc	(.74)	.79	(.70)
NWF-CLS	.78	.81	.79	Retell	(.68)	(.67)	(.65)
NWF-WWR	(.74)	.78	.79	Retell-Q	na	na	na
ORF-WRC	–	**.92**	**.92**	Daze	.80	.79	.78
ORF-Acc	–	**.91**	.89	**Grade 5**			
Retell	–	.79	.76	ORF-WRC	.85	.83	.83
Retell-Q	–	.75	(.74)	ORF-Acc	.76	(.73)	(.75)
Grade 2				Retell	(.71)	(.71)	(.69)
NWF-CLS	.79	–	–	Retell-Q	na	na	na
NWF-WWR	.79	–	–	Daze	.79	(.74)	.77
ORF-WRC	.82	.86	.86	**Grade 6**			
ORF-Acc	.80	.80	.80	ORF-WRC	.80	.80	.79
Retell	(.74)	(.73)	(.72)	ORF-Acc	(.67)	(.72)	(.71)
Retell-Q	na	na	na	Retell	(.73)	(.71)	(.71)
				Retell-Q	na	na	na
				Daze	.79	.75	.77

Note. AUC = Area Under the Curve; Beg = Beginning of Year; FSF = First Sound Fluency; LNF = Letter Naming Fluency; PSF = Phoneme Segmentation Fluency; NWF-CLS = Nonsense Word Fluency, Correct Letter Sounds Score; NWF-WWR = Nonsense Word Fluency, Whole Words Read Score; ORF-WRC = Oral Reading Fluency, Words Read Correctly Score; ORF-Acc = Oral Reading Fluency, Accuracy Score; Retell = Retell Fluency; Retell-Q = Retell Quality of Response Ranking; Daze = DIBELS Maze Adjusted Score.

Based on the results of the AUC analysis, we have established recommended goals for LNF, FSF, NWF-CLS, ORF, and DIBELS Maze (Daze). We also provide a goal and endorse the use of NWF-WWR at the end of Grade K (although NWF-WWR did not previously have a benchmark goal for that grade and time of year). We do not provide a recommended goal for PSF at either the end of kindergarten or at the beginning of Grade 1. Phonemic awareness skills are an

essential part of early literacy instruction; however, the value of PSF as a tool for universal screening beyond kindergarten is not supported in our research results. Future research is needed to determine if other types of phonemic awareness tasks may add value to the other DIBELS Next measures that are in place for grade K.

Lastly, we do not provide benchmark goals for the Retell measures. In part, this decision is because the AUC is greater than .75 in only three out of 17 occasions (middle and end of Grade 1, and beginning of Grade 3). We determined that Retell didn't perform consistently enough in the AUC analysis, nor did it reliably add value to predicting reading outcomes on the SAT10 in our earlier analysis (see Technical Brief Part I, UO CTL, 2012a), to warrant its *required* use for universal screening.

Even with our recommendation about the use of the Retell measure for screening, please keep in mind that our findings about Retell are focused on its use as a measure in the way it is defined in DIBELS Next. We acknowledge that reading comprehension is multi-faceted, and that *reading instruction should include elements of factual recall* (see sample Curriculum Map for Grade 3 reading instruction). However, the DIBELS Next Retell measure requires only a very small amount of factual recall, and is more dependent on expressive ability of a child than on a child's deep understanding of a passage. Researchers have long been searching for indicators of reading comprehension that can function better than Curriculum-Based Measures of Oral Reading (e.g., ORF) or Maze (e.g., Daze) tasks, but the search continues. For now, our best understanding of the data suggests that the DIBELS Next Retell measure is perfectly acceptable to use in certain situations, with certain students for whom you believe it necessary. However, we do not believe that requiring the administration of Retell for *all* students at each benchmark testing session is supported by research.

Changes in DMG former benchmark goals for recommended DIBELS Next measures. For *all* measures and time points where we recommend new benchmark goals, the recommended goal is more ambitious than the former goal. We have selected three measure by time of year and grade combinations to report here to illustrate the magnitude of these differences and the statistics associated with each one (see the DIBELS Next Recommended Goals document (UO CTL, 2012b) for a list of all new, UO recommended goals).

> ***NWF-CLS, middle of kindergarten***. We selected this example first, because it illustrates the *most typical* degree of discrepancy between the former and the UO recommended goals.
>
> ***ORF, beginning of Grade 3***. We selected this example because it represents the *least* degree of discrepancy between the former and the UO recommended goals.
>
> ***ORF, beginning of Grade 6***. We selected this example because it represents the *greatest* degree of discrepancy between the former and the UO recommended goals.

The magnitude of these differences are illustrated in Table 2 and represented visually in Figure 2. When viewing Table 2, note the following: (a) after the column headings, the first two rows of data reflect the recommended and the former *benchmark goals*; (b) the last two rows of data reflect the recommended and the former *cut points for risk*.

Also, for each of the three examples and each of the two types of goals, we provide five values: (a) the AUC; (b) the goal itself; (c) the sensitivity values and the corresponding specificity values listed in parentheses; (d) the SAT10 median percentile rank; and (e) the DDS percentile rank. For all *benchmark goals*, we report the median (or average) SAT10 percentile rank that is associated with students who score within four points of the goal. For example, in Grade 3, the median SAT10 percentile rank that is associated with students who start the year with an ORF score between 93 and 101 is the 48th percentile.

For all *cut points for risk*, we report the median SAT10 percentile rank that is associated with students who score below the cut point for risk. For example, in Grade 6, the median SAT10 percentile rank that is associated with students who start the year with an ORF score that is below 128 is the 20th percentile. The last column under each of the three examples is the exact percentile rank that is associated with the exact benchmark goal or cut score that is listed in the table. For example, in Grade 3, the DDS percentile rank that is associated with the beginning of year ORF benchmark goal of 97 is the 71st.

Table 2

Illustration of Changes in Benchmark Goals and Cut Points for Risk Using the Recommended Benchmark Goals Sample

	Kinder NWF-CLS Middle				Grade 3 ORF Beginning				Grade 6 ORF Beginning			
	Goal	*Sens.* *(Spec.)*	SAT10 *Mdn*	DDS PR	Goal	*Sens.* *(Spec.)*	SAT10 *Mdn*	DDS PR	Goal	*Sens.* *(Spec.)*	SAT10 *Mdn*	DDS PR
Benchmark Goals												
	AUC = .80				*AUC* = .84				*AUC* = .80			
Recommended (.90 sensitivity)	34	.90 (.47)	53rd	77th	97	.91 (.51)	48th	71st	150	.90 (.50)	56th	72nd
Former	17	.43 (.92)	31st	33rd	70	.65 (.83)	39th	39th	107	.34 (.95)	21st	21st
Cut Points for Risk												
	AUC = .78				*AUC* = .84				*AUC* = .80			
Recommended (.80 sensitivity)	25	.80 (.58)	24th	57th	73	.80 (.70)	14th	43rd	128	.80 (.65)	20th	49th
Former	8	.24 (.96)	10th	14th	55	.55 (.90)	8th	25th	90	.32 (.97)	6th	10th

Note. Sens. (Spec.) = Sensitivity and Specificity values; SAT10 *Mdn* = median (i.e., average) end-of-year SAT10 Total Reading percentile rank for a particular cut score or group of students; DDS PR = DIBELS Data System percentile rank; AUC = Area under the Receiver Operating Characteristic (ROC) Curve (.75 is minimum standard); Recommended = new DIBELS Next goals based on research at the Center on Teaching and Learning (.90 sensitivity); Former = original goals associated with DIBELS Next.

For all cases, note the discrepancy between the SAT10 percentile rank and the DDS percentile rank (i.e., in all cases, the SAT10 percentile rank is less than the DDS percentile rank). This distinction is a critical one because it highlights the overall risk level of our sample. Across all

grades, more than half of our sample is below the 33rd percentile on the SAT10. In Grade 2, half of our sample is below the 29th percentile; in kindergarten half of our sample is below the 36th percentile. *If we were to set a goal based on having no more than 40% of our student population score below that particular goal, then the goal level of performance in many cases would correspond to a national standard that was below the 25th percentile* (i.e., below the average range). We believe it is important to have a benchmark goal that is indicative of a performance standard that is valid and valuable across the country. For us, that goal is equal to performance that is at or above the 40th percentile on the SAT10.

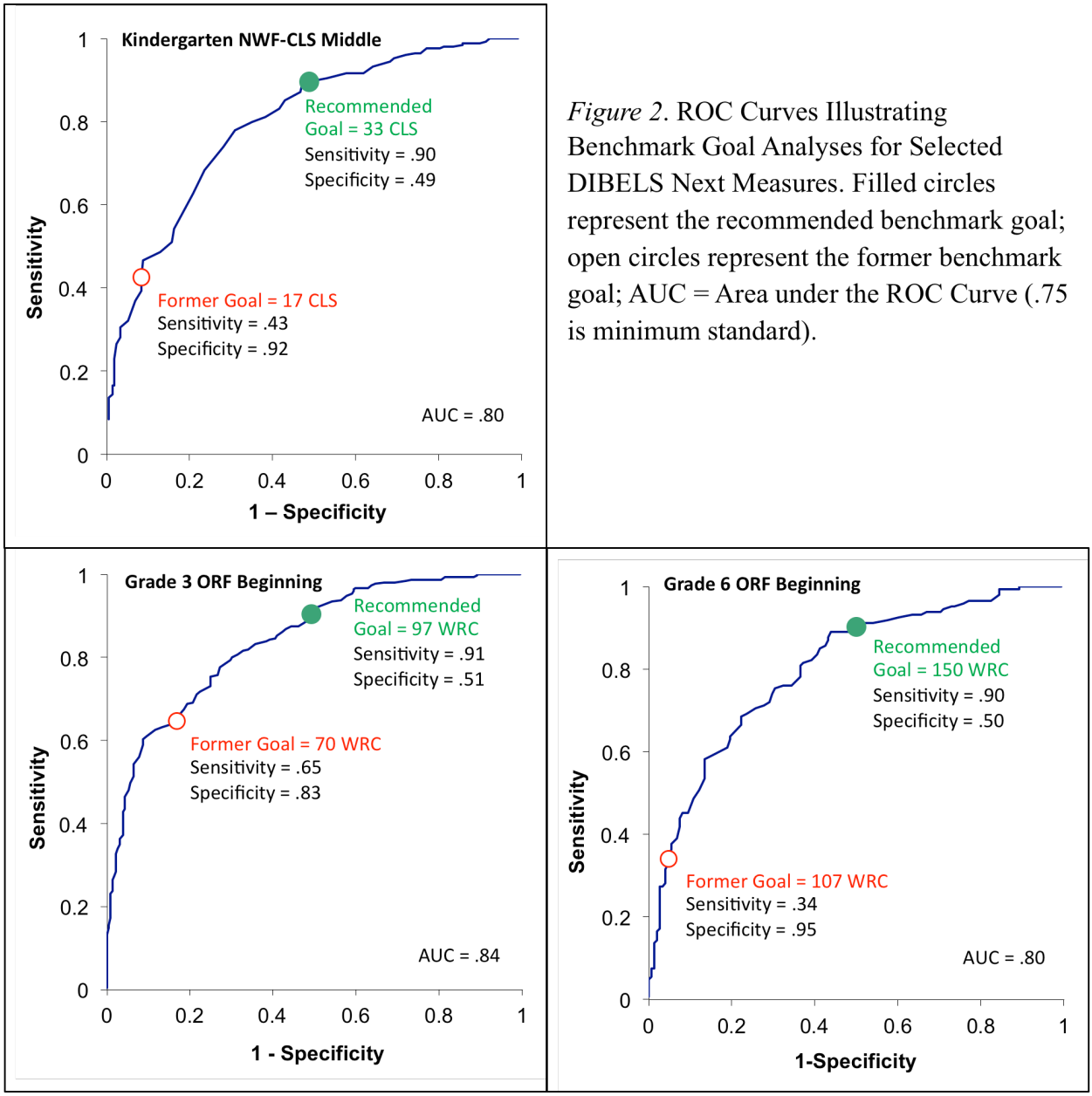

Figure 2. ROC Curves Illustrating Benchmark Goal Analyses for Selected DIBELS Next Measures. Filled circles represent the recommended benchmark goal; open circles represent the former benchmark goal; AUC = Area under the ROC Curve (.75 is minimum standard).

Conclusions

We listened to teachers and administrators who raised concerns about the large numbers of students who met benchmarks on the DMG former goals, yet failed to pass the standards on their high-stakes tests. This substantive and repeated feedback, coupled with the technical review of the DMG former goals, led to our own analysis of the DIBELS Next measures, in which we determined that the DMG former goals were not viable and that new *recommended goals* were required.

The measurement and research methodology used to set the former DMG DIBELS Next benchmark goals does not meet minimum standards for educational and psychological testing. By linking only the composite score to the outcome measure (and failing to link goals on the individual DIBELS Next measures to a comprehensive, end-of-year outcome measure), errors of prediction are compounded and estimates of goal-level performance are not consistent either within or across grades. The sample used in the DMG former goals analysis is not a representative sample, and makes comparisons for schools in many parts of the country inequitable. Finally, when test statistics that are base-rate sensitive are used to set national cut points, the results will be biased and fail to generalize.

The methodology used to set the DMG former DIBELS Next benchmark goals identified uniformly less-demanding scores for both the benchmark goal and cut point for risk than did the process used to determine the UO recommended goals. This feature of the DMG former goals has important implications for the schools that use the DIBELS Data System and the students they serve. For instance, *the accuracy in identifying the truly at-risk students (sensitivity) for the DMG former goals is consistently low: below .60 for two-thirds of the measure by time point combinations, with a range from .25 to .81*. Therefore, for two thirds of the assessments, students who meet the DMG former benchmark goals have less than a 60% chance of scoring at or above the 40th percentile on the SAT10 Total Reading Composite. In other words, these students will have failed to reach a healthy reading outcome, even when performance on screening measures indicated otherwise.

We acknowledge that the UO recommended goals will likely identify more students as needing additional instructional support. Our prevention-oriented approach to education services forms the basis for why, from the perspective of providing students the services they need to perform well academically, this change will be beneficial. Most students who need some level of support will get it and the provision of early intervening services to students is one of the most powerful tools educators have to promote long-term academic success. Of course, a critical variable in this process is teacher judgment. Teachers and school teams must use their professional knowledge and expertise to determine how best to serve their students, including determining the level of support students need to meet ambitious reading and learning outcomes. We believe that using the UO *recommended goals for* DIBELS Next measures will allow schools to more accurately predict students' future reading performance and intervene with confidence, knowing they are providing support to those students who need it.

References

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

Dynamic Measurement Group. (2011). *DIBELS Next Technical Manual*. Retrieved from https://dibels.org

Dynamic Measurement Group (2012, August 7). *Frequently asked questions about DIBELS Next: Why are the sixth grade benchmark goals lower than the fifth grade goals?* [web page]. Retrieved from http://dibels.org/faqsNext.html

Joint Committee on Testing Practices (2004). *Code of Fair Testing Practices in Education*. Washington, DC: Author. Retrieved from http://www.theaaceonline.com/codefair.pdf

Marzban, C. (2004). The ROC curve and the area under it as a performance measure. *Weather and Forecasting*, *19*, 1106-1114.

National Center for Education Statistics. (2011*). Common core of data: Public elementary/secondary school universe survey* [Data file and code book]. Retrieved from http://nces.ed.gov/ccd/pubschuniv.asp

National Center on Response to Intervention. (2011a). *Screening tools chart* [Results of the fourth annual review of screening tools]. Washington, DC: Author. Retrieved from http://www.rti4success.org/screeningTools

National Center on Response to Intervention. (2011b). *Standard protocol for evaluating response to intervention tools: Screening reading and math* [Screening tool evaluation protocol]. Washington, DC: Author.

O'Connell, A.A., & McCoach, D.B. (2008). *Multilevel modeling of educational data*. Charlotte, NC: Information Age Publishing, Inc.

Pearson Education, Inc. (2007). *Stanford Achievement Test—10th Edition (SAT10): Normative update.* Upper Saddle River, New Jersey: Author.

Peng, C.-Y. J., & So, T.-S. H. (2002). Logistic regression analysis and reporting: A primer. *Understanding Statistics*, *1*, 31-70.

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

Swets, J. A. (1996). *Signal detection theory and ROC analysis in psychology and diagnostics: Collected papers*. Hillsdale, NJ: Lawrence Erlbaum Associates.

University of Oregon, Center on Teaching and Learning (2012a). *2012-2013 DIBELS Data System Update Part I: DIBELS Next Composite Score*. (Technical Brief No. 1202). Retrieved from https://dibels.uoregon.edu/resources/DDS2012TechnicalBriefPart1.pdf

University of Oregon, Center on Teaching and Learning (2012b). *DIBELS Next Recommended Benchmark Goals*. Retrieved from https://dibels.uoregon.edu/docs/DIBELSNextRecommendedBenchmarkGoals.pdf

Recommended Citation:

University of Oregon Center on Teaching and Learning (2012). *2012-2013 DIBELS Data System Update Part II: DIBELS Next Benchmark Goals (Technical Brief No. 1203)*. Eugene, OR: University of Oregon.