

Developing Benchmark Goals and Cut-points for Risk: Odds of Achieving Subsequent Reading Goals

*Pacific Coast Research Conference
Coronado, CA
February 4, 2011*

*Roland H. Good III
Dynamic Measurement Group, Inc.
University of Oregon*

*Kelly Powell-Smith
Ruth A. Kaminski
Dynamic Measurement Group, Inc..*

February 4, 2011

Coronado, CA

1

Overview

- Panel:

The decision utility of educational assessment is enhanced when there are interpretive rules and recommendations for support associated with levels of performance. Benchmark goals provide a level of progress where students are judged to be making adequate progress. A cut point for risk indicates a level where a teacher can be confident the student will need additional intervention and support in order to make adequate progress. This panel will present alternative approaches to establishing benchmark goals and cut points for risk in the context of infants and toddlers, preschool students, and elementary students in kindergarten through sixth grade.

- Section:

One approach will examine the odds or likelihood of attaining important subsequent outcomes. The odds can be modeled with logistic regression and interpreted with receiver operator characteristic curves.

February 4, 2011

Coronado, CA

2

Benchmark Goal Study Research Questions

- The Benchmark Goal Study was designed to address three research questions:
 1. What levels of performance on *DIBELS Next* assessments predict a student is likely to score at or above the 40%ile on selected outcome measures?
 2. What levels of performance on *DIBELS Next* assessments predict a student is unlikely to score at or above the 40%ile on selected outcome measures?
 3. What are the correlations between *DIBELS Next* assessments and the Group Reading Assessment and Diagnostic Evaluation (GRADE), a criterion measure of reading proficiency that includes comprehension?

February 4, 2011

Coronado, CA

3

Participants

- Students recruited for the study were from 13 schools in five school districts representing five US regions.
- Participating school districts had a median of 10 years experience using DIBELS.
- Kindergarten through 6th grade students participated in *DIBELS Next* assessments ($n = 3,816$ total; 433 to 569 per grade). The percentage at benchmark ranged from 65% - 79% across grades and times of year.
- Subsamples of students participated in testing with an external criterion measure (Group Reading Assessment and Diagnostic Evaluation; GRADE) ($n = 1257$ total; 103 to 219 per grade). The GRADE subsample was 50% female on average across grades.

February 4, 2011

Coronado, CA

4

Participant Demographics

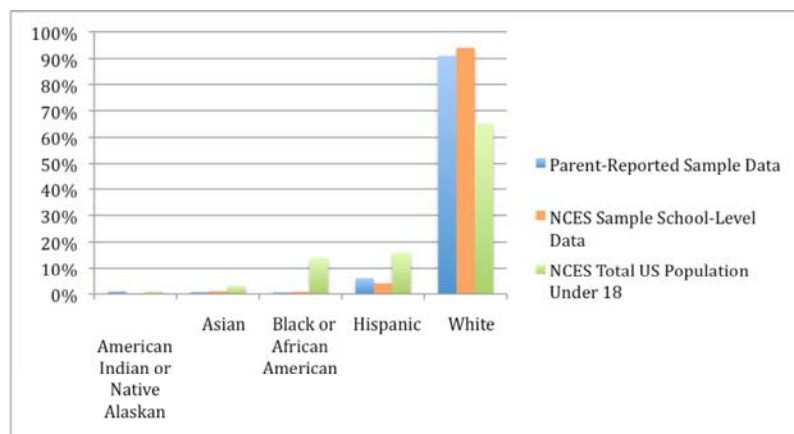


Figure 1. Racial/Ethnic Background

Participant Demographics

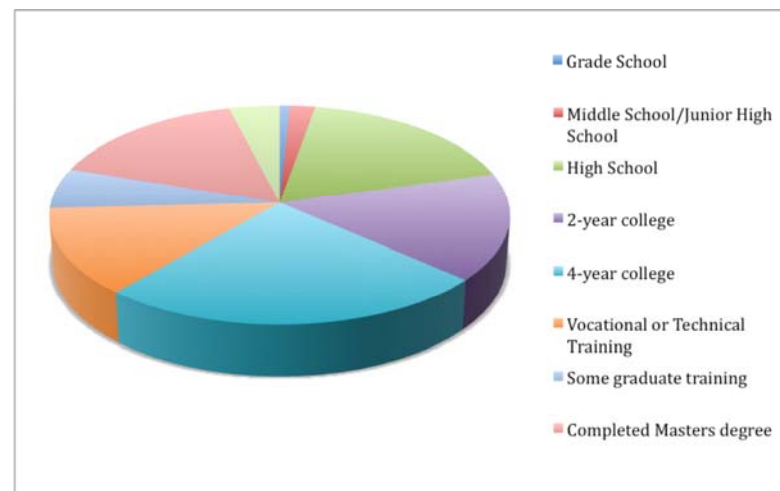


Figure 2: Parent-Reported Level of Education

Participant Demographics

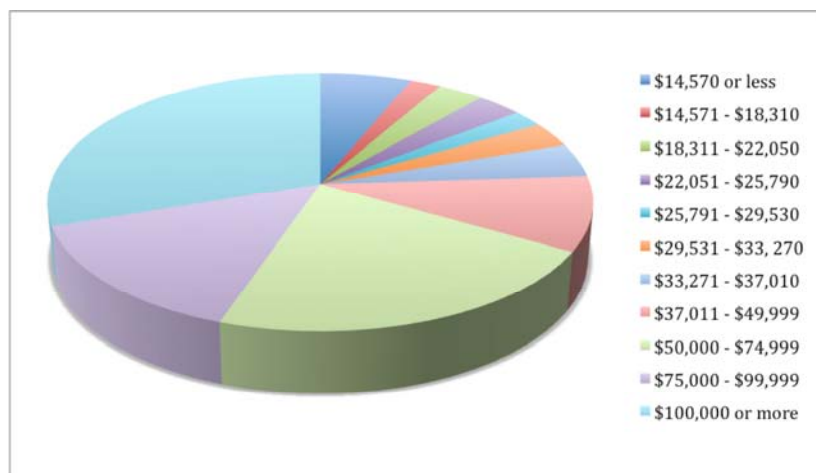


Figure 3: Parent-Reported Household Income

Measures: DIBELS Next

- The measures included all *DIBELS Next* assessments. *DIBELS Next* assessments include:
 - Letter Naming Fluency
 - First Sound Fluency
 - Phoneme Segmentation Fluency
 - Nonsense Word Fluency Correct Letter Sounds and Whole Words Read
 - Oral Reading Fluency Words Correct, Accuracy, and Retell.
 - Daze Adjusted Score (DIBELS-maze)
 - DIBELS Composite Score

Measures: Group Reading Assessment and Diagnostic Evaluation (GRADE)

- Un-timed and group administered. Appropriate for students in preschool through grade 12
- Five components and 16 subtests. Subtests combine to form the following composites:
 - Phonemic Awareness, Early Literacy Skills, Comprehension, Vocabulary, and Total Test.
 - We used the Total Test Raw Score for analyses.
- The GRADE has excellent reliability and validity for its intended purposes.
 - Reliability ranges from .77 to .98.
 - Correlation coefficients range from .69 to .86 with other group- and individually-administered achievement tests.

Procedures: Data Collection

- All Data were collected during the 2009-2010 school year
- *DIBELS Next* assessments were administered at regular benchmark intervals by trained school personnel using standardized procedures.
- GRADE testing occurred in the spring at the end of the year and was conducted across two to three sessions. Total testing time ranged from 60 to 90 minutes. The GRADE was administered by trained school personnel and onsite coordinators.

DIBELS Composite Score

- For each grade and time of year, the DIBELS Next measures that correlate highly with later outcomes are combined into a DIBELS Composite Score.
- Each measures is weighted so that **all contribute approximately equally** to the DIBELS Composite Score.
 - Weighted scores have approximately equal standard deviations.
- The DIBELS Composite Score represents a rich and broad sample of behavior.
- The DIBELS Composite Score conveys that *all of the aspects of reading proficiency are critical* – a student whose DIBELS Composite Score is At or Above Benchmark is reading **accurately**, at an **adequate rate**, and **attending to the meaning** of the passage.
- The DIBELS Composite is highly correlated with a broad range of reading outcomes.

For Example: Third Grade DIBELS Composite Score. Benchmark Goal: 220

Beginning, Middle and End of Year	
DORF Accuracy Percent	Accuracy Value
0% – 85%	0
86%	8
87%	16
88%	24
89%	32
90%	40
91%	48
92%	56
93%	64
94%	72
95%	80
96%	88
97%	96
98%	104
99%	112
100%	120

Beginning of Year Benchmark

DORF Words Correct = 99 [1]

Retell Score 39 x 2 = 78 [2]

Daze Adjusted Score 16 x 4 = 64 [3]

DORF Accuracy Percent: 98 %
100 x (Words Correct / (Words Correct + Errors))

Accuracy Value from Table = 104 [4]

DIBELS Composite Score (add values 1–4) = 345

Do not calculate the composite score if any of the values are missing (if DORF is below 40 and Retell is not administered, use 0 for Retell only on these worksheets).

Odds of being on track on the Middle of Year DIBELS Composite Score based on the Beginning of Year DIBELS Composite Score

Measure	Score Level	Score
DIBELS Composite Score	At or Above Benchmark	93%
	Below Benchmark	47%
	Well Below Benchmark	4%

DIBELS Composite Score is the Best Indicator of Reading Proficiency

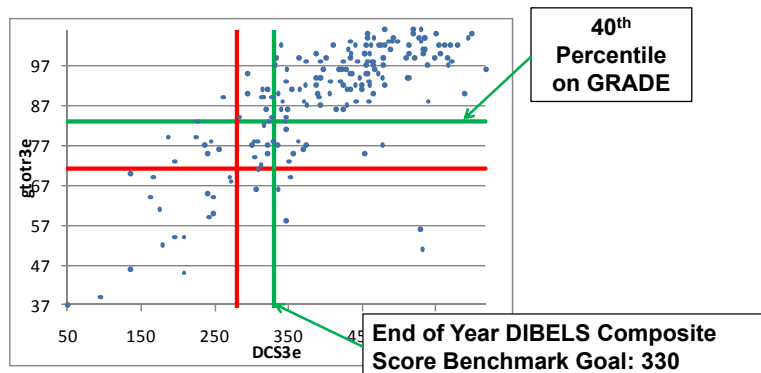
- DIBELS ORF is very good, DIBELS Composite Score is substantially better.
 - For example, beginning of year 3rd grade DORF Words Correct correlates with end of the year GRADE Total Score .66, which is very good.
 - Beginning of year 3rd grade DIBELS Composite Score correlates .73, explaining 10% more variance than DORF alone.
- DIBELS Composite Score beats the single best DIBELS Next measure at almost every grade and time of year.
- DIBELS Composite Score provides a more complete sample of reading behavior than any single measure.

Predictive Validity of the DIBELS Composite Score Compared to the Best Single DIBELS Measure by Grade and Time of Year

	Beginning of year		Middle of year		End of year	
	Best single measure	DIBELS Composite Score	Best single measure	DIBELS Composite Score	Best single measure	DIBELS Composite Score
Kindergarten						
GRADE Total raw score	0.52	0.50	0.47	0.48	0.40	0.37
End of year DIBELS Composite Score	0.43	0.52	0.65	0.71		
First grade						
GRADE Total raw score	0.43	0.55	0.64	0.70	0.75	0.77
End of year DIBELS Composite Score	0.71	0.73	0.83	0.89		
Second grade						
GRADE Total raw score	0.69	0.75	0.76	0.80	0.73	0.75
End of year DIBELS Composite Score	0.81	0.81	0.87	0.89		
Third grade						
GRADE Total raw score	0.66	0.73	0.67	0.78	0.66	0.75
End of year DIBELS Composite Score	0.86	0.88	0.86	0.90		
Fourth grade						
GRADE Total raw score	0.76	0.80	0.76	0.80	0.75	0.80
End of year DIBELS Composite Score	0.86	0.89	0.87	0.90		
Fifth grade						
GRADE Total raw score	0.69	0.76	0.64	0.76	0.66	0.77
End of year DIBELS Composite Score	0.84	0.86	0.86	0.90		
Sixth grade						
GRADE Total raw score	0.64	0.71	0.59	0.68	0.61	0.73
End of year DIBELS Composite Score	0.86	0.90	0.87	0.91		

Third Grade DIBELS Composite Score for End of Year (DCS3e) and GRADE Total Raw Score (gtotr3e)

- Group Reading Assessment and Diagnostic Evaluation (GRADE) is a high quality, nationally norm-referenced test.
- .75 correlation to GRADE Total Raw Score at End of Year



What is the Purpose of Benchmark Goals and Screening for Risk in Education?

Different standards, procedures, and requirements are necessary if our purpose is:

- To quickly *identify* students that are likely to need additional support to *prevent* later academic difficulty.
 - To accurately identify students who are the *true* Tier 3 students or who have a *true* learning disability early.
- To specify important and meaningful future goals—a level of skills at a point in time where we can change the odds to being in favor of an individual's meeting subsequent goals.

We are troubled by the purpose of identifying *true* Tier 3 students. We think the future is not set. Tier 3 is not a characteristic of the student. There are no *true* Tier 3 students. Tier 3 is a level of support necessary for the student to make adequate progress. No fate but what we make.

Our purpose is to prevent reading difficulty and enhance reading outcomes by providing targeted, differentiated instruction early.

Skill Level	Skill level definition	National (public school) percent of fourth grade students <i>scoring below</i> (pp. 16, 52-53)	Nation (public) percent of fourth grade students from diverse backgrounds <i>scoring below</i> (pp. 54 & 57)
Basic	Basic denotes partial mastery of prerequisite knowledge and skills that are fundamental for proficient work at a given grade.	34%	54%, 51%, 49%, 50%
Proficient	Proficient represents solid academic performance. Students reaching this level have demonstrated competency over challenging subject matter.	68%	86%, 83%, 80%, 83%

Note: Students from diverse backgrounds includes students identified as Black, Hispanic, American Indian/Alaska Native, and eligible for free/reduced-price school lunch. From data reported in Lee, Grigg, & Donahue (2007).

17

Goal: Adequate Reading Skills

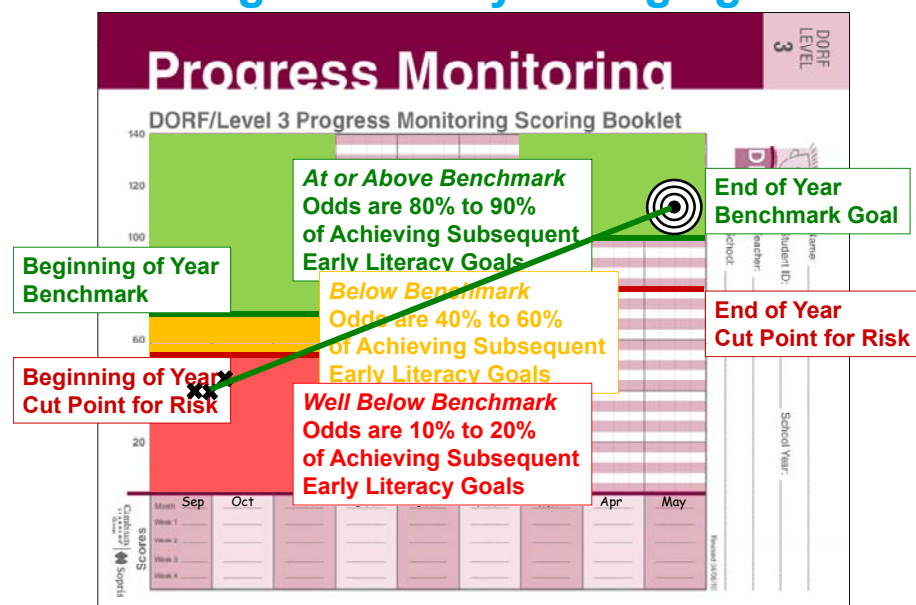
- Adequate reading skills should generalize across different state, national, and published reading tests.
- Adequate reading skills are not a normative decision, but are a socio-political judgment.
- The 40th percentile or above on a high quality, nationally norm-referenced test can serve as an approximation for adequate reading performance.
- Students at or above the 40th percentile on a high quality, nationally norm-referenced test are on track to be rated Basic or above on NAEP.
- We used the Group Reading Assessment and Diagnostic Evaluation (GRADE) in our initial research to provide an initial approximation of adequate reading skills.

February 4, 2011

Coronado, CA

18

Building Futures by Changing Odds



February 4, 2011

Coronado, CA

19

Evidence Base, Score Level, Likely Need for Support

Odds of achieving subsequent early literacy goals	Outcome: Looking back	Screening: Looking forward
	Score level	Likely need for support to achieve subsequent early literacy goals
80% to 90%	At or Above Benchmark scores at or above the benchmark goal	Likely to Need Core Support
40% to 60%	Below Benchmark scores below the benchmark goal and at or above the cut point for risk	Likely to Need Strategic Support
10% to 20%	Well Below Benchmark scores below the cut point for risk	Likely to Need Intensive Support

The fundamental rationale for benchmark goals and screening decisions is based on the odds of achieving subsequent early literacy goals.

February 4, 2011

Coronado, CA

20

Primary Design Specifications for DIBELS Goals and Cut Points for Risk

- **Primary Specification:** At or Above Benchmark Decision on initial (screening) DIBELS assessment should provide favorable odds (80% -- 90%) of achieving subsequent reading outcomes. Benchmark Goal should provide a level where we are reasonably confident the student is making adequate progress.
- Below Benchmark Decision on initial DIBELS assessment should provide 50 -- 50 odds (40% -- 60%) of achieving subsequent reading outcomes. Below the Benchmark Goal but above the Cut Point should provide a zone of uncertainty where we don't know if the student is making adequate progress or not.
- Well Below Benchmark Decision on initial DIBELS assessment should provide low odds (10% -- 20%) of achieving subsequent reading outcomes -- **unless** intensive intervention is implemented. Below the Cut Point should provide a zone where we are reasonably confident the student will not make adequate progress -- unless we provide additional support.

Secondary Specifications for Benchmark Goals and Cut Points

- Marginal percents for the predictor close to marginal percents for the outcome.
 - The sample for the Benchmark Goal Study was a relatively high performing sample.
 - We tried have them appear equally high performing on DIBELS Next and the GRADE.
- Logistic Regression Analysis
 - Logistic regression predicted odds of about 60% or better at the exact goal score.
 - Logistic regression predicted odds of about 40% or below at the exact cut point for risk score.

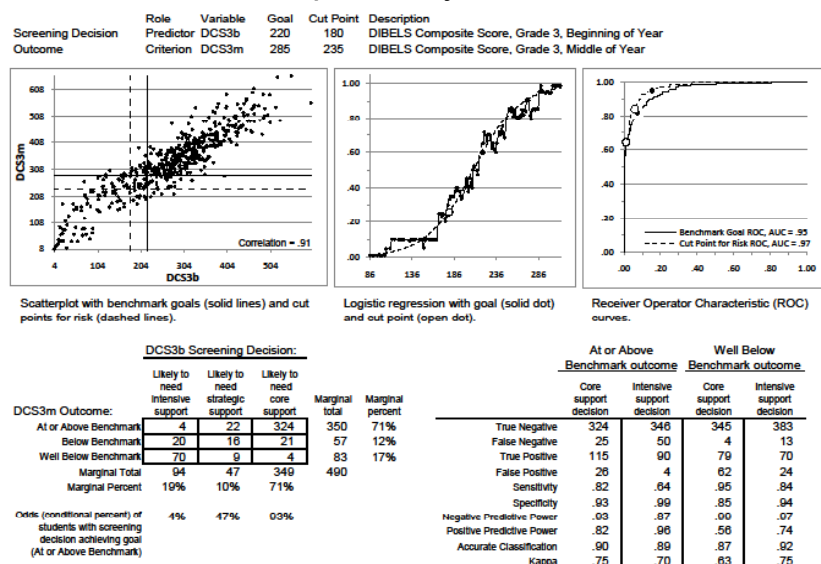
Other Considerations DIBELS Goals and Cut Points

- Other considerations
 - Receiver Operator Characteristic Curve (ROC) analysis with large area under curve
 - Other metrics for decision utility
 - sensitivity,
 - specificity,
 - percent correct classification,
 - kappa
 - Coherent pattern of goals across measures and grades.

Setting Benchmark Goals and Cut Points for Risk

1. Examine scatterplot illustrating the relation between the screening assessment (earlier assessment or predictor) and the outcome assessment (later assessment).
 - DIBELS is a step-by-step model, so the outcome of one step is the predictor of the next step.
2. Examine the table of counts for each zone of the scatterplot.
3. Primary: Consider odds of students with each screening decision achieving goal.
4. Secondary: Consider marginal percents
5. Secondary: Consider logistic regression analysis
6. Other: Consider ROC curve and decision utility metrics
7. Other: Consider the overall pattern of goals and cut points.

Example Analysis Detail

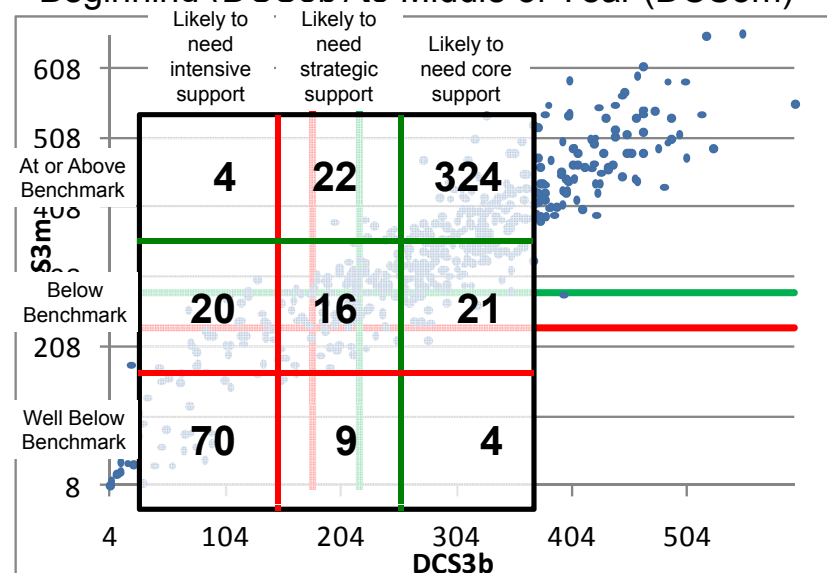


February 4, 2011

Coronado, CA

25

Third Grade DIBELS Composite Score for Beginning (DCS3b) to Middle of Year (DCS3m)



26

Third Grade DIBELS Composite Score for Beginning (DCS3b) to Middle of Year (DCS3m)

DCS3b Screening Decision:

DCS3m Outcome:	Likely to need intensive support	Likely to need strategic support	Likely to need core support	Marginal total
At or Above Benchmark	4	22	324	350
Below Benchmark	20	16	21	57
Well Below Benchmark	70	9	4	83
Marginal Total	94	47	349	490

- Primary consideration: Odds of achieving outcome goal.
- Secondary consideration: Marginal Percents

February 4, 2011

Coronado, CA

27

Primary consideration: Odds of achieving goal

DCS3b Screening Decision:

DCS3m Outcome:	Likely to need intensive support	Likely to need strategic support	Likely to need core support	Marginal total
At or Above Benchmark	4	22	324	350
Below Benchmark	20	16	21	57
Well Below Benchmark	70	9	4	83
Marginal Total	94	47	349	490

- Core support beginning of year screening decision: 324 of 349 students achieve the middle of year goal, or 93% odds.
- Strategic support: 22 of 47 students achieve the goal, or 47% odds.
- Intensive support: 4 of 94 students achieve the goal, or 4% odds.

February 4, 2011

Coronado, CA

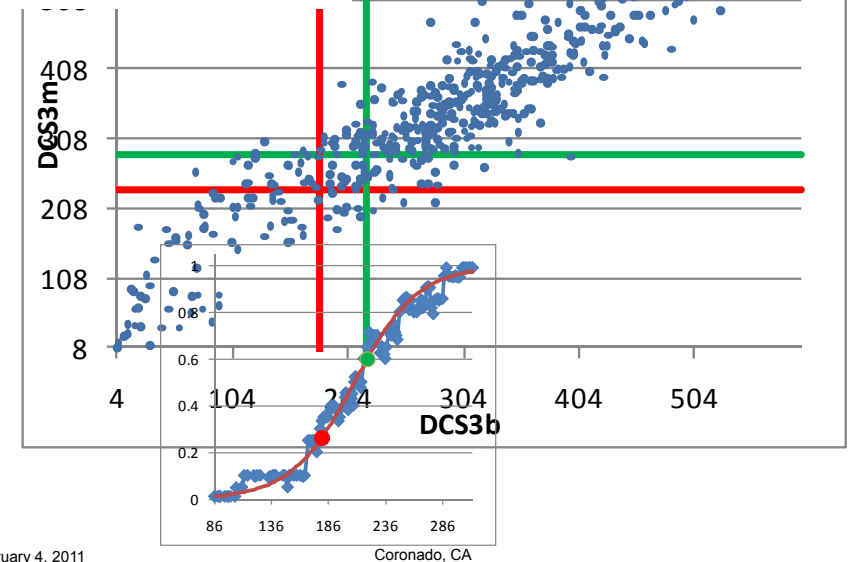
28

Also Considered Marginal Percents

DCS3b Screening Decision:					
DCS3m Outcome:	Intensive support decision	Strategic support decision	Core support decision	Marginal total	Marginal percent
At or Above Benchmark	4	22	324	350	71%
Below Benchmark	20	16	21	57	12%
Well Below Benchmark	70	9	4	83	17%
Marginal Total	94	47	349	490	
Marginal Percent	19%	10%	71%		

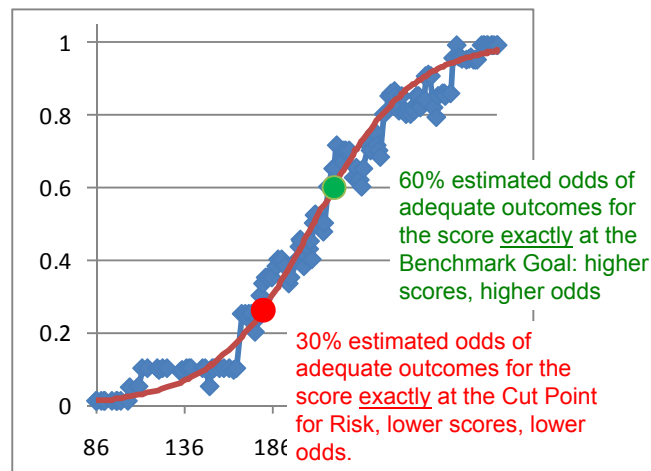
- Percent At or Above Benchmark at beginning of year is very close to the percent At or Above Benchmark in the middle of the year.
- Desirable for the screening decision to identify about the same percent of students that are expected on the outcome.

Moving Odds: Logistic Regression



Logistic Regression Estimates Odds of Adequate Outcomes for each Score

- Blue diamonds are moving proportion with adequate outcome.
- Red line is logistic regression estimated odds of adequate outcomes.

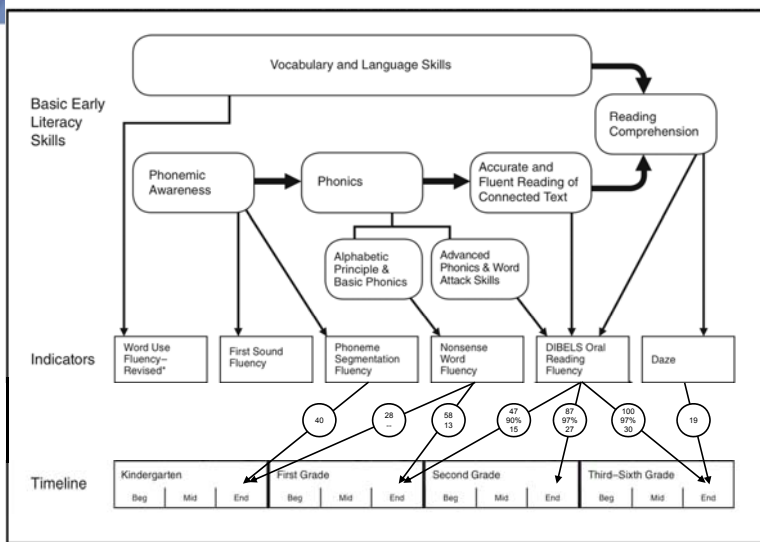


DIBELS is a Step-by-Step Model: Beginning to Middle; Middle to End;

- Mastering each step puts the odds in favor of mastering the next step.
 - At or Above Benchmark:** Odds are generally 80% to 90% of achieving subsequent benchmark goals and important reading outcomes. **Student is likely to make adequate progress with effective core instruction.**
 - Below Benchmark:** Odds are generally 40% to 60% of achieving subsequent benchmark goals and important reading outcomes. **Student is likely to need strategic support to make adequate progress.**
 - Well Below Benchmark:** Odds are generally 10% to 20% of achieving subsequent benchmark goals and important reading outcomes. **Student is likely to need intensive support to make adequate progress.**
- Contiguous Continuity. Each step is a continuous process with a strong linkage. Each step is contiguous with the next step.



End of Year Benchmark Goals



*Word Use Fluency—Revised (WUF-R) is available as an experimental measure from <http://dibels.org/>. For Nonsense Word Fluency, the first number is the Correct Letter Sounds goal, and the second number is the Whole Words Read goal. For DIBELS Oral Reading Fluency, the first number is the Words Correct goal, the second number is the Accuracy goal, and the third number is the Retell goal. For example, at the end of second grade a student should be able to read 87 words with 97% accuracy and retell of 27 words relevant to the passage. Third grade benchmark goals are illustrated; benchmark goals for grades 4 through 6 are available.

33

DIBELS® Next: Summary of Benchmark Goals and Cut Points for Risk

DIBELS Composite Score	26	122	119	113	130	155	141	190	238	220	285	330	290	330	391	357	372	415	344	358	380
First Sound Fluency (FSF)	10	30																			
Letter Naming Fluency (LNF)	5	20																			
Phoneme Segmentation Fluency (PSF)	20	40	40																		
Nonsense Word Fluency (NWF)	17	28	27	43	58	54															
DIBELS Oral Reading Fluency (DORF)	23	47	52	72	87	70	86	100	90	103	115	111	120	130	107	109	120				
Words Correct	16	32	37	55	65	55	68	80	70	79	95	96	101	105	90	92	95				
Accuracy	78%	90%	90%	96%	97%	95%	96%	97%	96%	97%	98%	98%	98%	99%	97%	97%	98%				
Retell	15	16	21	27	20	26	30	27	30	33	33	36	36	27	29	32					
Retell Quality of Response	8	13	18	10	18	20	14	20	24	22	25	25	16	18	24						
Daze	8	11	19	15	17	24	18	20	24	18	14	14	15								
Grade	Kindergarten	First Grade	Second Grade	Third Grade	Fourth Grade	Fifth Grade	Sixth Grade														

This is a summary of the DIBELS Next benchmark goals. For a full description, see the DIBELS Next Benchmark Goals and Composite Score document available from <http://dibels.org/>. DIBELS is a registered trademark of Dynamic Measurement Group, Inc.

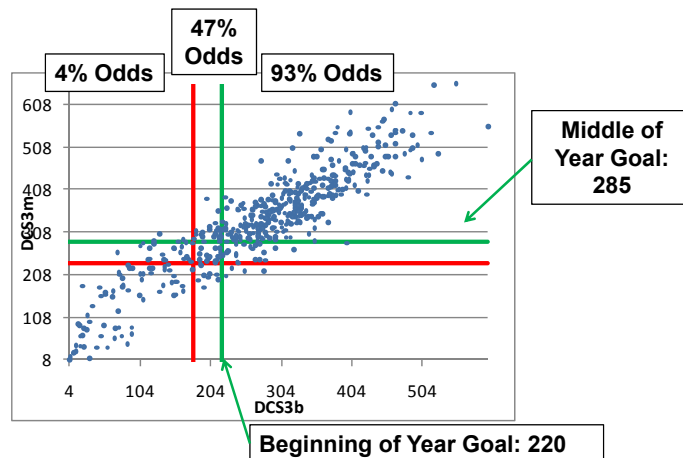
February 4, 2011

Coronado, CA

34

Third Grade DIBELS Composite Score for Beginning of Year (DCS3b) and Middle of Year (DCS3m)

- .91 correlation to DIBELS Composite Score at Middle of Year



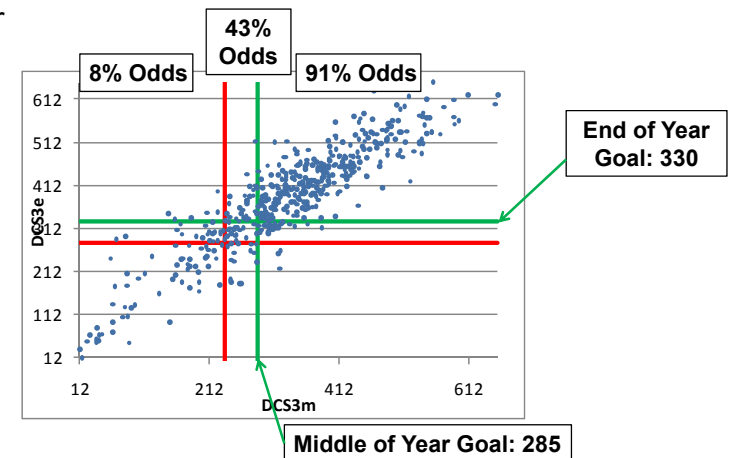
February 4, 2011

Coronado, CA

35

Third Grade DIBELS Composite Score for Middle of Year (DCS3m) and End of Year (DCS3e)

- .90 correlation to DIBELS Composite Score at End of Year



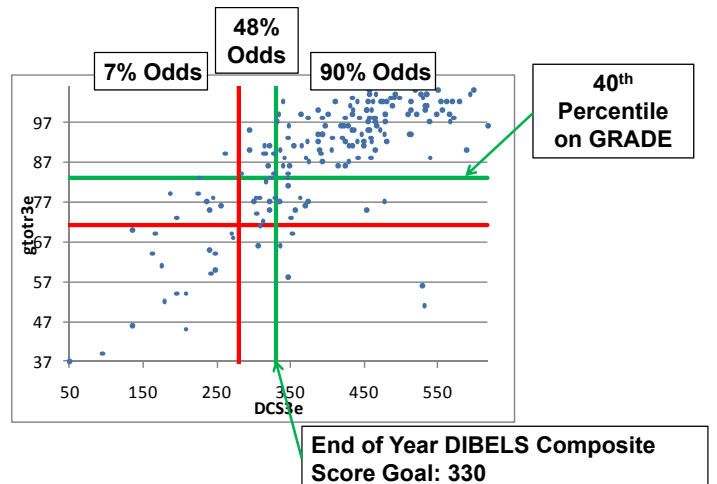
February 4, 2011

Coronado, CA

36

Third Grade DIBELS Composite Score for End of Year (DCS3e) and GRADE Total Raw Score (gtotr3e)

- .75 correlation to GRADE Total Raw Score at End of Year



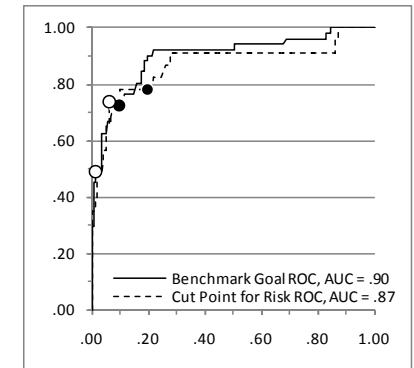
February 4, 2011

Coronado, CA

37

Receiver Operator Characteristic Curve

- Larger area under the curve indicates favorable trade off of sensitivity and specificity.
- Decision points in the upper left bend of the curve indicate a favorable balance of sensitivity and specificity.



Receiver Operator Characteristic (ROC) curves.

February 4, 2011

Coronado, CA

38

Other Decision Utility Metrics End of Third Grade

We are troubled by the terminology. We think a "True Positive" is actually a student for whom we were not effective in ruining the prediction. A "False Positive" is a student for whom we have changed the future.

	At or Above Benchmark outcome		Well Below Benchmark outcome	
	Core support decision	Intensive support decision	Core support decision	Intensive support decision
True Negative	123	134	132	154
False Negative	14	26	5	6
True Positive	37	25	18	17
False Positive	13	2	32	10
Sensitivity	.73	.49	.78	.74
Specificity	.90	.99	.80	.94
Negative Predictive Power	.90	.84	.96	.96
Positive Predictive Power	.74	.93	.36	.63
Accurate Classification	.86	.85	.80	.91
Kappa	.63	.56	.39	.63

Role	Variable	Goal	Cut Point	Description
Screening Decision	Predictor DCS3e	330	280	DIBELS Composite Score, Grade 3, End of Year
Outcome	Criterion gtotr3e	83	71	GRADE Total Test, Grade 3, End of Year

February 4, 2011

Coronado, CA

39

Early Intervention and Prevention are Active Ingredients Between Screening and Outcomes

- The effectiveness of the school-wide system of instruction can change the odds.
 - Differences in the effectiveness of Tier 1 instruction and Tier 2 & 3 intervention change the underlying relation between screener and outcome.
 - Less effective school-wide system Tier 1 instruction can decrease the odds of achieving subsequent early literacy goals for students who are at or above benchmark.
 - Increasing the effectiveness of Tier 2 & 3 intervention can increase the odds of achieving subsequent early literacy goals for students who are at risk.

February 4, 2011

Coronado, CA

40

Building Futures

- Key Point: The student's outcome is unknown and not fixed at the time of the screening. Instead, the outcome is the result of the targeted, differentiated instruction and intervention we provide as a direct result of the screening information.
- **Our instructional goal is to ruin screening predictions**
- For Example: If a child screens as at high risk on a measure of early literacy skills in Kindergarten, we know they are likely to need additional instructional support to be successful. Their later outcome, their reading skills in first grade for example, are a direct result of the targeted, differentiated instruction and early intervention that we provide.