

DIBELS® Benchmark Goals

DIBELS Summit
 Santa Ana Pueblo, NM
 February 18, 2009

Kelli D. Cummings
Dynamic Measurement Group, Inc.

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

Rachael Latimer & Maya E. O'Neil
Dynamic Measurement Group, Inc.

Acknowledgements

- Wireless Generation
- DIBELS Beta 1 Research Study Partners
 - *Annie Hommel MA, Research Assistant and Site Coordinator*
- Dynamic Measurement Group Data Analysis Team
 - *Rachael Latimer, Research Assistant*
 - *Maya O'Neil MS, Assistant Data Analyst*
 - *Josh Wallin, Director of Operations and Technology Manager*

Overview

- Importance of evaluating screening tools in an educational context.
- Two commonly used metrics, Sensitivity and Specificity, are problematic in an educational context because they assume:
 - A true, dichotomous outcome.
 - A gold standard of the outcome that is generally agreed upon.
 - No intervening active ingredient between screening and outcome...they are affected by differences in quality of instruction.
- Our recommendation:
 - Use likelihood or odds of achieving important educational outcomes given the screening decision to evaluate screening assessments.

What about meaningful rates of progress? What about generally effective instruction?

Evidentiary Considerations for the Educational Decisions Required for Response to Intervention Models

Educational Decision	Evidentiary Consideration		
	Reliable	Normative Context	Meaningful
1. Is the student making adequate year-to-year progress?	X	X	X
2. Is the student receiving generally effective instruction?	?	?	?
3. Is the student making adequate week-to-week progress?	+/-	+/-	+/-

Note. X = generally strong and persuasive evidence. ? = level of evidence is unestablished. +/- = emerging evidence base.

Sample from mClass Data System

- Data were gathered from 8890 schools in 1226 districts across 50 states for students who were in first grade in the 2004-2005 academic year and were followed longitudinally into their second grade year in the 2005-2006 academic year.
- All data were collected using the Palm® version of DIBELS.
- Participating school districts received training on DIBELS and the Palm during implementation.
- All data were collected using district procedures, district trained and supervised data collectors.

Descriptive Stats for mClass Samples

	mClass samples					Monte Carlo study		
	Full mClass Sample	500 random sub-sample	137 district sub-sample	District 1	District 2	137 district sample	District 1	District 2
<i>n</i>	58811	500	46154	490	466	46154	490	466
ORF Gr 2 EOY								
Mean	91.93	91.85	91.09	61.87	84.16	90.92	71.56	79.08
<i>sd</i>	37.11	37.26	37.51	35.58	34.32	38.30	35.59	34.32
NWF Gr 1 EOY								
Mean	62.87	62.80	62.04	46.10	52.29	62.03	46.11	52.30
<i>sd</i>	30.56	29.64	31.05	29.56	26.04	31.05	29.54	26.06
correlation	.63	.65	.63	.59	.62	.68	.64	.61

- 500 random sample from the full data set is for illustrative purposes.
- 137 district sample has complete data for at least 100 students in each district.
- A Monte Carlo study was conducted to model the 137 districts in the mClass sample with bivariate normal random data with (a) the same correlation as the full mClass sample, (b) the same NWF mean, NWF standard deviation, and ORF standard deviation as each district, (c) but with the ORF district mean set to be the same number of standard deviation units from the full mClass sample mean as the NWF district mean.

Overview

- Importance of evaluating screening tools in an educational context.
- Two commonly used metrics, Sensitivity and Specificity, are problematic in an educational context because they assume:
 - A true, dichotomous outcome.
 - A gold standard of the outcome that is generally agreed upon.
 - No intervening active ingredient between screening and outcome...they are affected by quality of instruction.
- Our recommendation:
 - Use likelihood or odds of achieving important educational outcomes to evaluate screening assessments.

Purpose of Screening Tools in Education

- To quickly identify the likelihood that a student will need additional help to *prevent* a later academic difficulty.
- To specify important and meaningful goals—a point at which we change the odds to being in favor of an individual's meeting subsequent goals.
- Key Point:** Outcomes are unknown and are likely *not even present* at the time of the screening. Instead, outcomes eventuate or come into being as a result of the differentiated instruction and intervention provided as a direct result of the screening information.
- 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. The eventual outcome, their reading skills in first grade, for example, is a direct result of the differentiated instruction and intervention that are provided.

We need to critically evaluate our screening tools for educational decisions

- We need to evaluate the:
 - Reliability of the measures,
 - Validity of the measures,
 - Decision utility of the measures (*are the goals meaningful and important?*),
 - Consequential validity of the measures.
- Sensitivity and Specificity indices may not be the best metrics to evaluate educational screening measures.
- Sensitivity and specificity were developed for and are most appropriate when:
 - There is a true, dichotomous outcome.
 - There is a gold standard of the outcome that is generally agreed upon.
 - There is no intervening active ingredient. Only when there is no intervening active ingredient are the constructs of “False Positive” and “False Negative” even meaningful.
 - For example, a screening test for tuberculosis.

Overview

- Importance of evaluating screening tools in an educational context.
- Two commonly used metrics, Sensitivity and Specificity, are problematic in an educational context because they assume:
 - A true, dichotomous outcome.
 - A gold standard of the outcome that is generally agreed upon.
 - No intervening active ingredient between screening and outcome...they are affected by quality of instruction.
- Our recommendation:
 - Use likelihood or odds of achieving important educational outcomes to evaluate screening assessments.

For Example, Screening for Tuberculosis

Screening Decision:
Positive TB Negative TB

True State (Outcome):
Negative for tuberculosis

FP: False Positive	TN: True Negative
TP: True Positive	FN: False Negative


True State (Outcome):
Positive for tuberculosis

- Sensitivity:** Of individuals who truly have tuberculosis, what proportion are identified as having tuberculosis by the screening test?
- Specificity:** Of individuals who truly do not have tuberculosis, what proportion are identified as not having tuberculosis on the screening test?

$$\frac{TP}{TP + FN}$$

$$\frac{TN}{FP + TN}$$

Screening for Tuberculosis, Sensitivity and Specificity Make Sense

- There is a true state, and it is a dichotomous one (TB/not TB) not one of degree (a patient doesn't have a little bit of TB).
- A gold standard of the true state is generally agreed upon. We are able to know with reasonable certainty whether the person has TB or not.
-  Sensitivity and Specificity are used to evaluate the accuracy of the screening tool *before* treatment or action takes place. There is no active ingredient or treatment between screening and gold standard identification of the true state.

In an Educational Context, We Need More Sense Than Sensitivity

- To evaluate screening tools in education, our recommendation is to use the likelihood of achieving important educational outcomes because:
 - The outcome is continuous.
 - There is no general agreement on a specific assessment or cutpoint on the assessment that discriminates adequate and not adequate skills.
 - And especially because there is intervening instruction and intervention occurring between the screening assessment and the outcome. **When there is intervening instruction and intervention, the constructs of “False Positive” and “False Negative” are not meaningful.**

Screening for Adequate Reading Skills

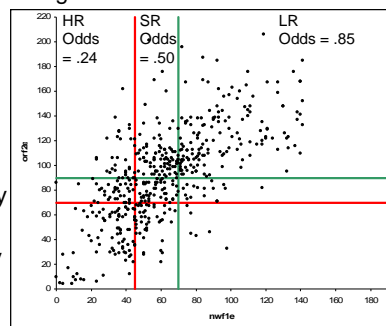
True State (Outcome):	Screening Decision:		
	High Risk	Some Risk	Low Risk
Adequate Reading skills (Negative for reading difficulty)	n_{11}	n_{12}	n_{13}
Uncertain Reading skills (We don't agree if adequate or not)	n_{21}	n_{22}	n_{23}
Poor Reading Skills (Positive for Reading Difficulty)	n_{31}	n_{32}	n_{33}

- Low Risk Likelihood or Odds:** Of individuals who are identified as low risk on the screening test, what proportion achieve adequate reading skills on the outcome assessment? $\frac{n_{13}}{n_{13} + n_{23} + n_{33}}$
- Some Risk Likelihood or Odds:** Of individuals who are identified as some risk on the screening test, what proportion achieve adequate reading skills on the outcome assessment? $\frac{n_{12}}{n_{12} + n_{22} + n_{32}}$
- High Risk Likelihood or Odds:** Of individuals who are identified as high risk on the screening test, what proportion achieve adequate reading skills on the outcome assessment? $\frac{n_{11}}{n_{11} + n_{21} + n_{31}}$

For Example, DIBELS Assessment

First Grade End of Year NWF Initial Assessment:

High Risk Some Risk Low Risk



Second End of Year ORF Outcome:
Low Risk Reading Fluency

Some Risk Reading Fluency

High Risk Reading Fluency

- Low Risk Likelihood or Odds:** Of students who are Low Risk on DIBELS NWF at end of first grade, 85% are Low Risk on end of second grade ORF.
- Some Risk Likelihood or Odds:** Of students who are Some Risk on DIBELS NWF at end of first, 50% are Low Risk on end of second grade ORF. We just don't know if they are on track or not.
- High Risk Likelihood or Odds:** Of students who are High Risk on DIBELS NWF at end of first, 24% are Low Risk on end of second grade ORF

We can impose a 2-by-2 Model on Reading Assessment, but it Doesn't Really Fit

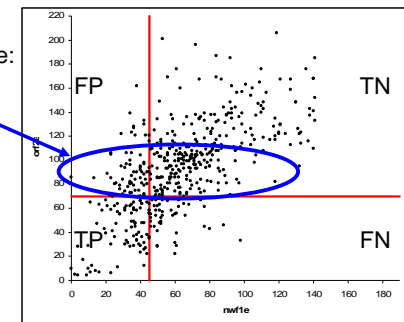
What about these students?

Second End ORF Outcome:
Not High Risk

High Risk

DIBELS Alphabetic Principle:

High Risk Not High Risk



- Sensitivity:** Of students **who truly have poor reading**, what proportion are identified as having poor reading by DIBELS?
- Specificity:** Of students **who truly do not have poor reading**, what proportion are identified as not having poor reading on DIBELS?

$$\frac{TP}{TP + FN}$$

$$\frac{TN}{FP + TN}$$

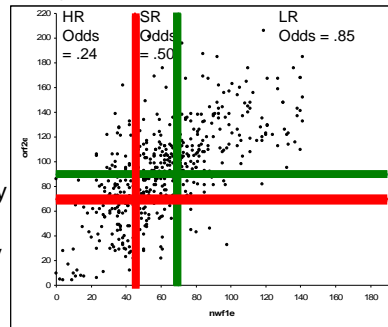
Educational Assessment is a Three-by-three World

First Grade End of Year NWF Initial Assessment:
High Risk Some Risk Low Risk

Second End of Year ORF Outcome:
Low Risk Reading Fluency

Some Risk Reading Fluency

High Risk Reading Fluency



- Using 2-by-2 logic in a 3-by-3 world, 4 different decisions must be evaluated:
 - LRD-LRO: Low Risk Screening Decision with a Low Risk Outcome.
 - LRD-HRO: Low Risk Screening Decision with a High Risk Outcome.
 - HRD-LRO: High Risk Screening Decision with a Low Risk Outcome
 - HRD-HRO: High Risk Screening Decision with a High Risk Outcome.**

Note: Odds based on Full WG sample, $n = 58811$. Scatterplot based on a random sub-sample of WG sample, $n = 500$.

February 18, 2008

DIBELS Summit, Albuquerque, NM

17

Using Sensitivity or Specificity to Evaluate or Compare Screening Tools is Meaningless

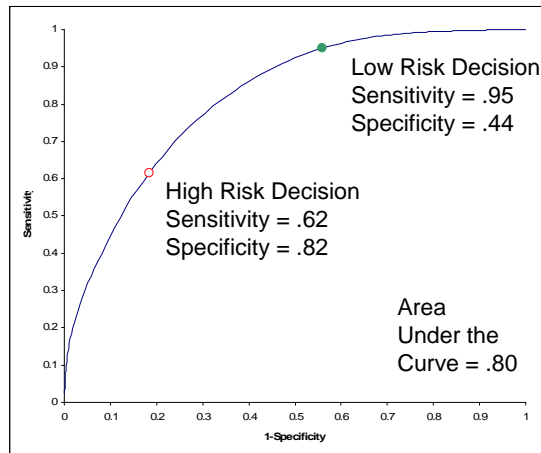
- It is meaningless to compare sensitivity indices on different tests (Swets, 1988) because:
 - Sensitivity *depends on the cutpoint for risk* that is selected. As we increase the cutpoint, sensitivity increases,
 - But, there is a trade-off. As we increase the cutpoint, the specificity decreases.
 - Area under the Receiver Operator Characteristic (ROC) Curve is the only general index of the accuracy of a screening measure that is independent of the cutpoint selected.
 - However, the ROC curve *also* depends on having a gold standard of the outcome criterion. For tuberculosis, this is not a problem. For reading skills in an educational context, as we have seen, this is a significant problem.
 - At the very least, we need separate ROC curves for high risk outcomes and low risk outcomes.

February 18, 2008

DIBELS Summit, Albuquerque, NM

18

ROC Curve for Second Grade, End of Year ORF Low Risk Outcome



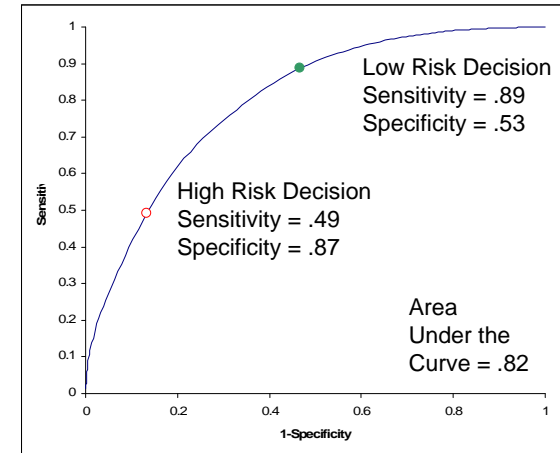
Full WG Sample, $n = 58811$

February 18, 2008

DIBELS Summit, Albuquerque, NM

19

ROC Curve for Second Grade, End of Year ORF High Risk Outcome



Full WG Sample, $n = 58811$

February 18, 2008

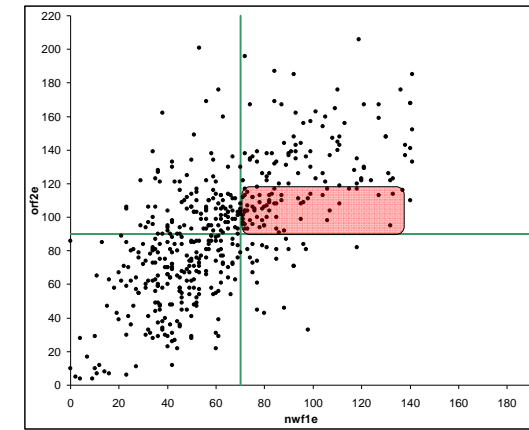
DIBELS Summit, Albuquerque, NM

20

Overview

- Importance of evaluating screening tools in an educational context.
- Two commonly used metrics, Sensitivity and Specificity, are problematic in an educational context because they assume:
 - A true, dichotomous outcome.
 - A gold standard of the outcome that is generally agreed upon.
 - No intervening active ingredient between screening and outcome...they are affected by quality of instruction.
- Our recommendation:
 - Use likelihood or odds of achieving important educational outcomes to evaluate screening assessments.

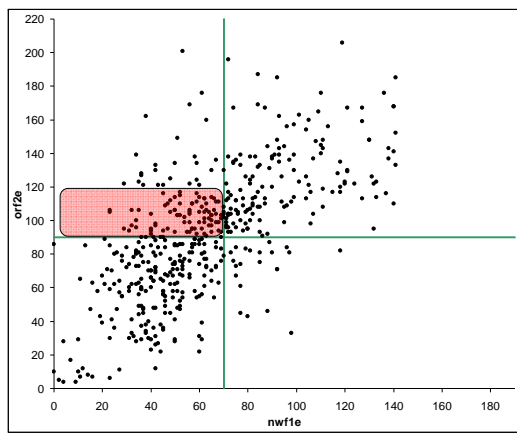
Educational Decision Making Also Has the Problem of Differential Tier 1 Effectiveness



More Effective Tier 1 Instruction:
As more students who screened low risk achieve the outcome, specificity and sensitivity increase.

- In a context with a greater (lesser) Tier 1 Instructional Effectiveness, more (less) students who screened negative will be negative on the outcome.
- The underlying relation between screener and outcome is changed, because selected students would move vertically on the scatterplot.

Educational Decision Making Also Has the Problem of Differential Tier 2 & 3 Effectiveness



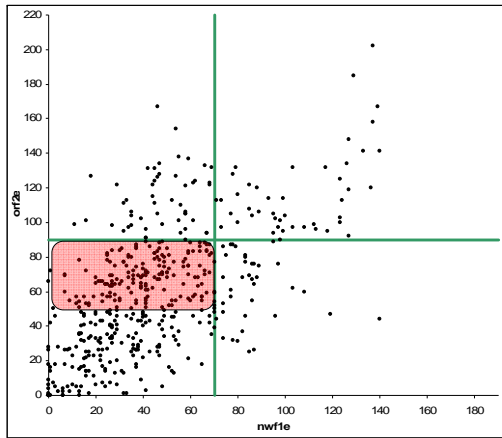
More Effective Tier 2 & 3 Intervention :
As more students who screened high or some risk achieve the outcome, specificity and sensitivity decrease.

- In a context with a greater (lesser) Tier 2 & 3 Instructional Effectiveness, fewer (more) students who screened positive will be positive on the outcome.
- Again, the underlying relation between screener and outcome is changed, because selected students would move vertically on the scatterplot.

The Big Ideas

- Differences in the effectiveness of Tier 1 instruction and Tier 2 & 3 intervention change the underlying relation between screener and outcome.
- Increasing the effectiveness of Tier 1 instruction **increases** measures of sensitivity and specificity.
- Increasing the effectiveness of Tier 2 & 3 intervention **decreases** measures of sensitivity and specificity.
- Increasing the effectiveness of the schoolwide system (Tier 1, 2, and 3 support) results in chaotic, unpredictable, and uninterpretable changes in measures of sensitivity and specificity.

Sensitivity & Specificity Logic Doesn't Work



Sample District 1	
Decision Baserate	0.82
True Negative	45
False Negative	45
True Positive	349
False Positive	51
Sensitivity	0.89
Specificity	0.47
Negative Predictive Power	0.50
Positive Predictive Power	0.87
Accurate Classification	0.80

- Consider Sample District 1.
- Are we really comfortable saying these students are “True Positives”? Or are they failures of our Tier 2 and Tier 3 intervention?

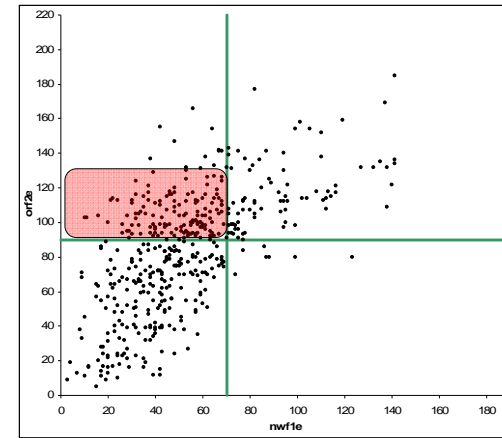
February 18, 2008

DIBELS Summit, Albuquerque, NM

Note. Outcome baserate would be .80.

25

Sensitivity & Specificity Logic Doesn't Work



Sample District 2	
Decision Baserate	0.81
True Negative	82
False Negative	7
True Positive	223
False Positive	154
Sensitivity	0.97
Specificity	0.35
Negative Predictive Power	0.92
Positive Predictive Power	0.59
Accurate Classification	0.65

- In Sample District 2, students with similar initial skills are achieving adequate reading skills. Does this mean they are “False Positives”? Or are they successes of our Tier 2 and Tier 3 intervention?

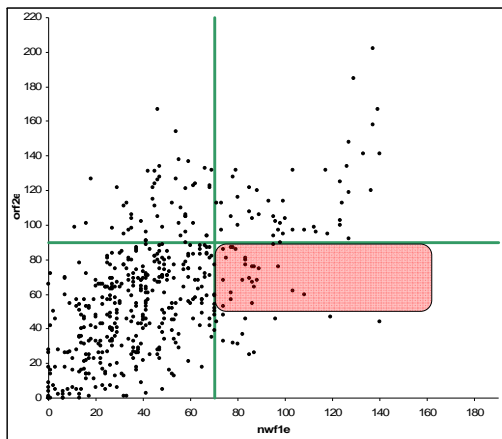
February 18, 2008

DIBELS Summit, Albuquerque, NM

Note. Outcome baserate would be .49.

26

Sensitivity & Specificity Logic Doesn't Work



Sample District 1	
Decision Baserate	0.82
True Negative	45
False Negative	45
True Positive	349
False Positive	51
Sensitivity	0.89
Specificity	0.47
Negative Predictive Power	0.50
Positive Predictive Power	0.87
Accurate Classification	0.80

- Consider Sample District 1 again.
- Do we really want to consider these students to be “False Negatives”? Or are they failures of our Tier 1 instruction?

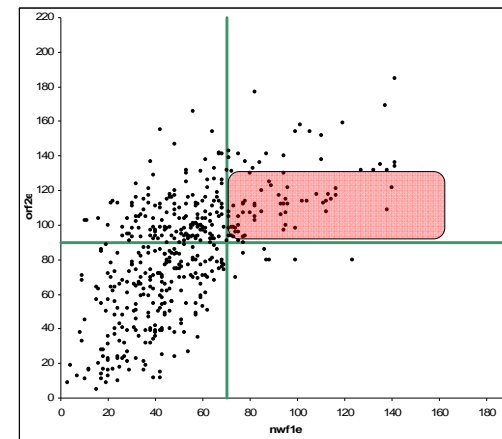
February 18, 2008

DIBELS Summit, Albuquerque, NM

Note. Outcome baserate would be .80.

27

Sensitivity & Specificity Logic Doesn't Work



Sample District 2	
Decision Baserate	0.81
True Negative	82
False Negative	7
True Positive	223
False Positive	154
Sensitivity	0.97
Specificity	0.35
Negative Predictive Power	0.92
Positive Predictive Power	0.59
Accurate Classification	0.65

- In Sample District 2, students with similar initial skills are almost all achieving adequate reading skills. Does this mean they are “True Negatives”? Or are they successes of our Tier 1 instruction?
- **A fundamental problem is that outcomes are not set, fixed, immutable, “true” at the time of screening. Instead, outcomes are achieved by instruction and intervention.**

February 18, 2008

DIBELS Summit, Albuquerque, NM

Note. Outcome baserate would be .49.

28

Overview

- Importance of evaluating screening tools in an educational context.
- Two commonly used metrics, Sensitivity and Specificity, are problematic in an educational context because they assume:
 - A true, dichotomous outcome.
 - A gold standard of the outcome that is generally agreed upon.
 - No intervening active ingredient between screening and outcome...they are affected by quality of instruction.
- Our recommendation:
 - Use likelihood or odds of achieving important educational outcomes to evaluate screening assessments.

Design Specifications of DIBELS Cutpoints

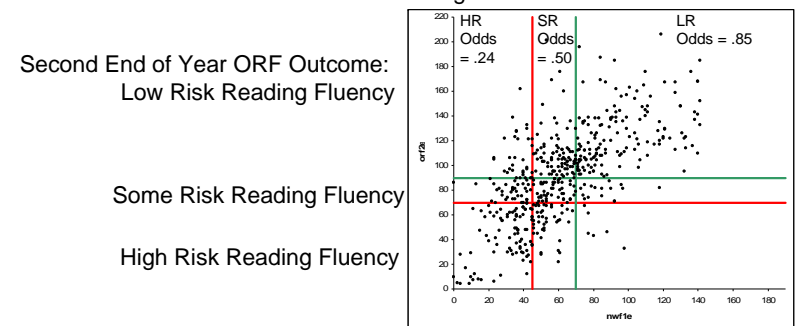
- **Primary Specification:** Low Risk Decision on initial DIBELS assessment should result in the favorable likelihood, or odds, (85% +/- 5%) of achieving subsequent reading health outcomes. In other words, a zone where we are reasonably confident the student has adequate skills.
- Some Risk Decision on initial DIBELS assessment should result in 50 – 50 odds (50% +/- 5%) of achieving subsequent reading health outcomes. In other words, a zone of uncertainty where we don't know if the student is on track or not.
- High Risk Decision on initial DIBELS assessment should result in low odds (15% +/- 5%) of achieving subsequent reading health outcomes – unless intensive intervention is implemented. In other words, a zone where we are reasonably confident the student does not have adequate skills.

Linking Screening Decisions to Instruction: The Purpose is to Improve Outcomes

- Likelihood or odds are a proxy for what it would take to change outcomes. What would it take to ruin the prediction?
- **Low Risk:** odds are in favor of achieving subsequent outcomes.
 - Likely to be easier to teach.
 - Likely to need good Tier 1 instruction (no guarantees!).
- **Some Risk:** means we don't know the likely outcome. If we do nothing special, the odds are 50 – 50. Maybe we should do something to improve the odds?
 - Likely to be harder to teach.
 - Likely to require more resources for success.
 - Likely to require more effective, intensive instruction.
 - Likely to need additional Tier 2 support.
- **High Risk:** means the odds are against achieving adequate outcomes – unless we provide intensive intervention.
 - Likely to be much harder to teach.
 - Likely to require even more resources for success.
 - Likely to require more extremely careful, effective, intensive intervention.
 - Likely to need effective Tier 3 intervention.

High Risk, Some Risk, and Low Risk Decisions

First Grade End of Year NWF Initial Assessment:
High Risk Some Risk Low Risk



- High risk, some risk, and low risk likelihood of outcomes (odds) vary with instructional context in interpretable ways.

Note: Odds based on Full WG sample, $n = 58811$. Scatterplot based on a random sub-sample of WG sample, $n = 500$.

Decision Utility of DIBELS with the Full MClass Sample

Odds of Achieving ORF Benchmark Outcomes (Criterion)					
Initial Support Decision Based on First Grade EOY NWF (Screen)		G1 ORF EOY	G2 ORF BOY	G2 ORF MOY	G2 ORF EOY
	Low Risk >= 70	.92	.85	.91	.85
	Some Risk 45 - 69	.54	.49	.60	.50
	High Risk < 45	.22	.25	.31	.24
	N=	253375	177576	157548	58811

DIBELS Beta 1 Validation Study

- 19 elementary schools, from 6 school districts across the U.S.
- Included students in grades K – 6
- Schools were DIBELS users (range of experience 4 – 9 years) who volunteered to participate
- All schools were trained via webcast on new and substantially revised DIBELS measures (FSF, WUF-R, NWF)
- All schools agreed to collect DIBELS data and to record additional information as part of the study

Research Questions

- What are the range of scores on DIBELS® Next measures by grade and time of year?
- What are the intercorrelations among DIBELS® Next measures within grade and time of year?
- What are the predictive correlations among DIBELS® Next measures across the school year?
- What is the decision utility of the DIBELS benchmark goals and cut points?

NWF-Middle of Kindergarten

Descriptive Statistics for DIBELS Kindergarten Measures

Measure	Mean	SD	Min	25th	50th	75th	Max	N
Middle of year								
Nonsense Word Fluency	22.65	16.73	0	12	20	30	143	2286
End of year								
Nonsense Word Fluency	41.64	23.18	0	26	38	52	143	2259

Note. 25th = 1st quartile; 50th = 2nd quartile; 75th = 3rd quartile. Correlation between Nonsense Word Fluency middle and end of year scores is .73*(2222); the number of subjects with pair-wise complete data is reported in parentheses. Correlations that explain greater than 20% of variability are denoted by an asterisk (*).

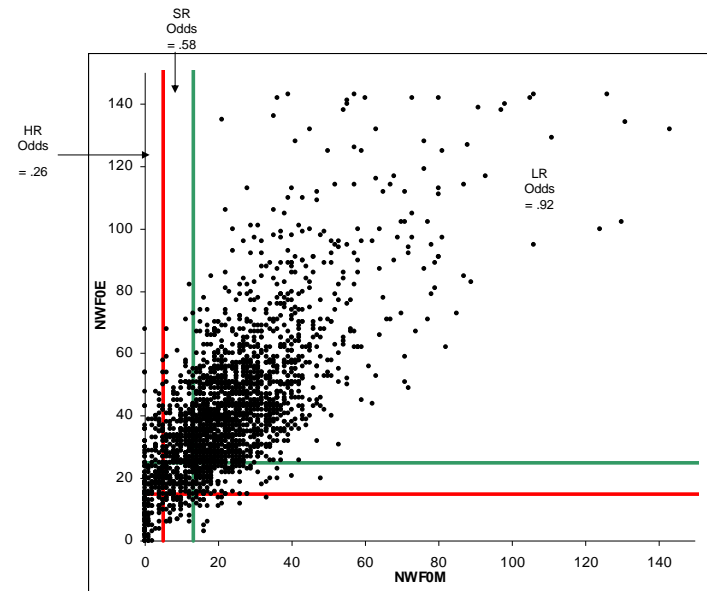
NWF-Middle of Kindergarten

Likelihood of Achieving NWF Kindergarten End of Year Benchmark Outcomes for Decisions Based on NWF Kindergarten Middle of Year Scores

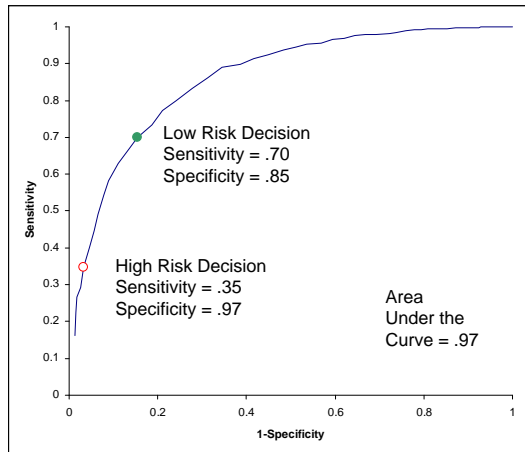
Likelihood of achieving benchmark outcomes	
Low Risk: NWF score is 13 or more	.92
Some Risk: NWF score is 5 to 12	.58
High Risk: NWF score is 0 to 4	.26
Area under the ROC curve	
Low risk score on outcome	.97
High risk score on outcome	.98

Note. Likelihood is reported as a conditional probability of a low risk outcome given NWF MOY score. NWF = Nonsense Word Fluency; MOY = Middle of Year; EOY = End of Year; ROC = Receiver Operator Characteristic.

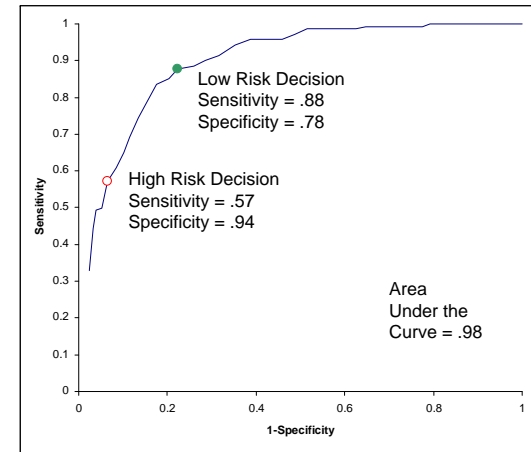
NWFOM to NWFEOE



ROC for Kindergarten, End of Year NWF Low Risk Outcome



ROC for Kindergarten, End of Year NWF High Risk Outcome



NWF-Beginning of 1st Grade

Table X
Descriptive Statistics for DIBELS First Grade Measures

Measure	Mean	SD	Min	25th	50th	75th	Max	N
Beginning of year								
Nonsense Word Fluency	38.11	25.51	0	22	32	47	143	2138
Middle of year								
Nonsense Word Fluency	60.50	30.01	0	40	53	73	143	2149
Oral Reading Fluency	37.33	30.94	0	14	28	51	161	1005
End of year								
Nonsense Word Fluency	81.83	34.45	0	55	76	108	143	2135
Oral Reading Fluency	75.40	40.40	0	45	71	102	219	2133

Note. 25th = 1st quartile; 50th = 2nd quartile; 75th = 3rd quartile.

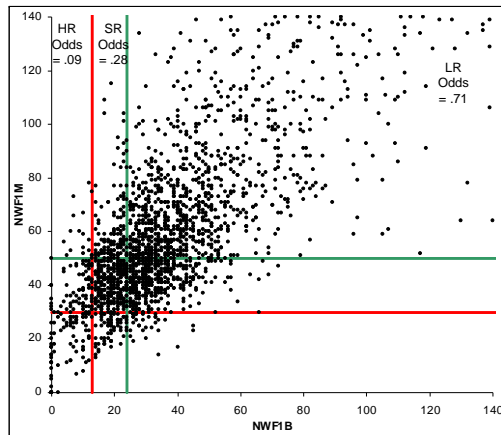
NWF-Beginning of 1st Grade

Table X
Likelihood of Achieving Benchmark Outcomes for Decisions Based on NWF First Grade Beginning of Year Scores

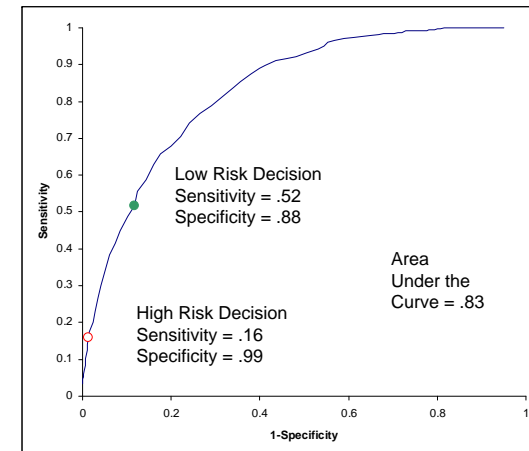
	NWF MOY	ORF MOY	NWF EOY	ORF EOY
Likelihood of achieving benchmark outcomes				
Low Risk: NWF score is 24 or more	.71	.84	.69	.93
Some Risk: NWF score is 13 to 23	.28	.37	.34	.64
High Risk: NWF score is 0 to 12	.09	.09	.13	.24
Area under the ROC curve				
Low risk score on outcome	.83	.86	.79	.87
High risk score on outcome	.87	.93	.81	.92

Note. Likelihood is reported as a conditional probability of a low risk outcome given NWF BOY score. NWF = Nonsense Word Fluency; ORF = Oral Reading Fluency; BOY = Beginning of Year; MOY = Middle of Year; EOY = End of Year; ROC = Receiver Operator Characteristic.

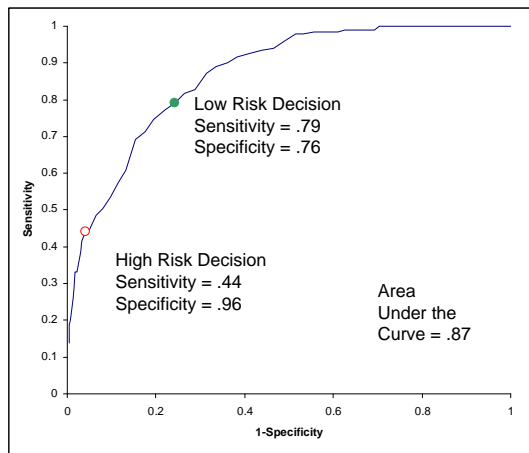
NWF-Beginning of 1st Grade



ROC for First Grade, Middle of Year NWF Low Risk Outcome



ROC for First Grade, Middle of Year NWF High Risk Outcome

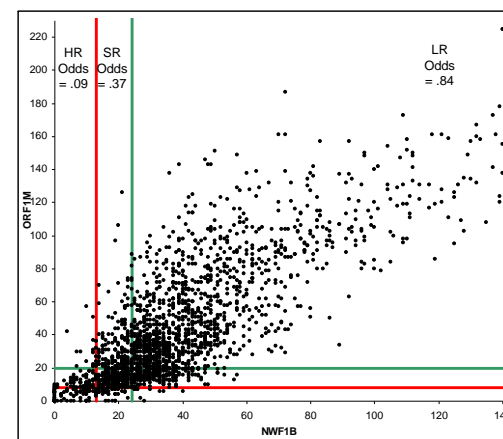


February 18, 2008

DIBELS Summit, Albuquerque, NM

45

NWF-Beginning of 1st Grade

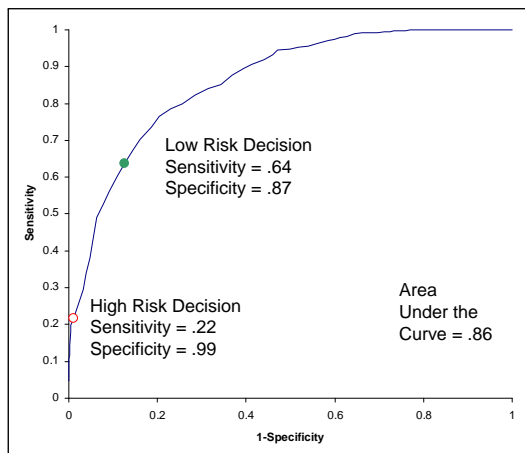


February 18, 2008

DIBELS Summit, Albuquerque, NM

46

ROC for First Grade, Middle of Year ORF Low Risk Outcome

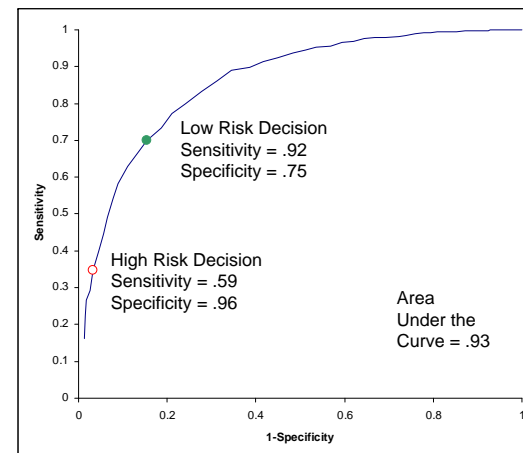


February 18, 2008

DIBELS Summit, Albuquerque, NM

47

ROC for First Grade, Middle of Year ORF High Risk Outcome

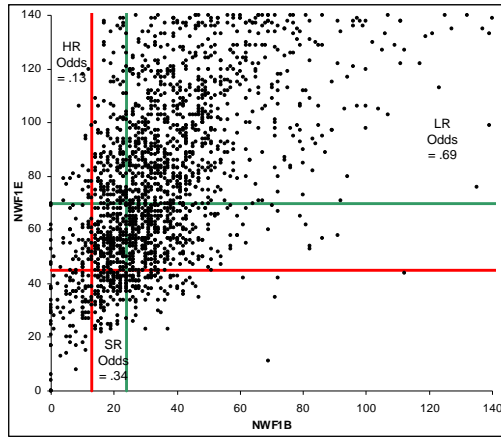


February 18, 2008

DIBELS Summit, Albuquerque, NM

48

NWF1B to NWF1E

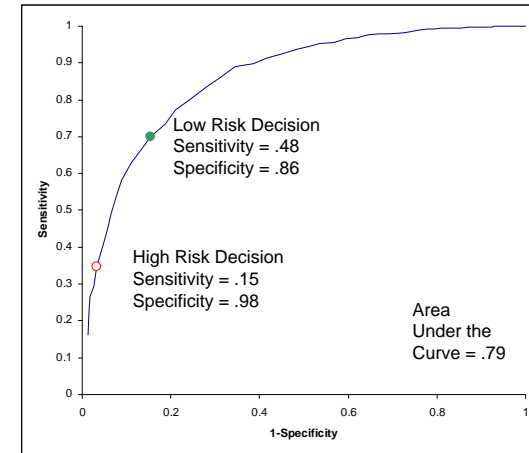


February 18, 2008

DIBELS Summit, Albuquerque, NM

49

ROC for First Grade, End of Year ORF Low Risk Outcome

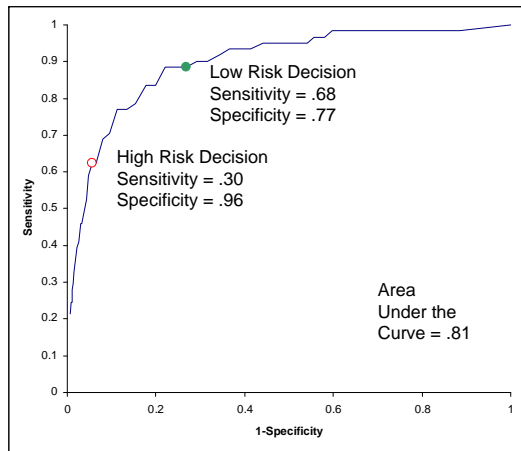


February 18, 2008

DIBELS Summit, Albuquerque, NM

50

ROC for First Grade, End of Year ORF High Risk Outcome

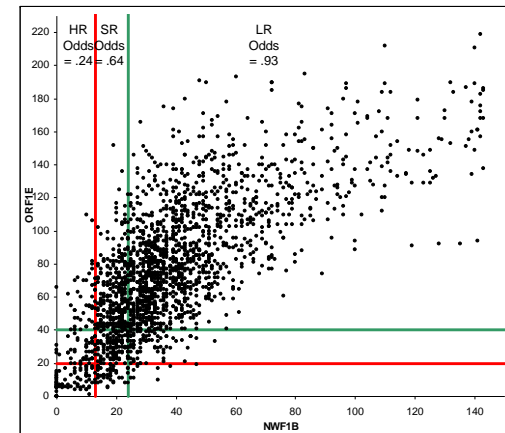


February 18, 2008

DIBELS Summit, Albuquerque, NM

51

NWF1B to ORF1E

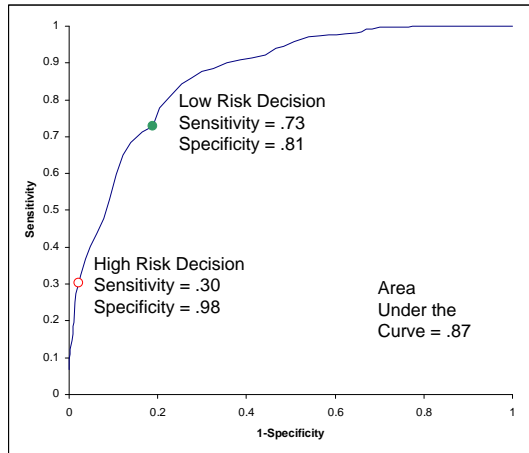


February 18, 2008

DIBELS Summit, Albuquerque, NM

52

ROC for First Grade, End of Year ORF Low Risk Outcome

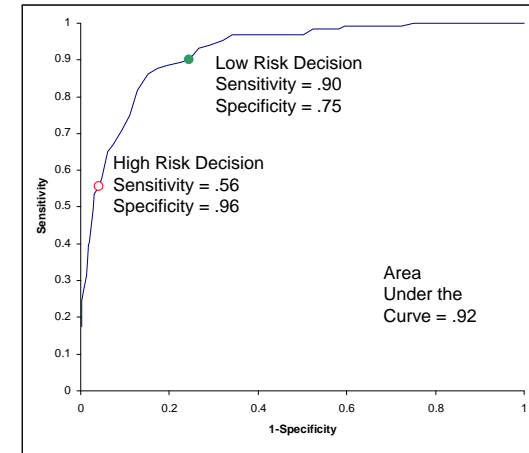


February 18, 2008

DIBELS Summit, Albuquerque, NM

53

ROC for First Grade, End of Year ORF High Risk Outcome



February 18, 2008

DIBELS Summit, Albuquerque, NM

54

NWF-Middle of 1st Grade

Table X

Descriptive Statistics for DIBELS First Grade Measures

Measure	Mean	SD	Min	25th	50th	75th	Max	N
Middle of year								
Nonsense Word Fluency	60.50	30.01	0	40	53	73	143	2149
Oral Reading Fluency	37.33	30.94	0	14	28	51	161	1005
End of year								
Nonsense Word Fluency	81.83	34.45	0	55	76	108	143	2135
Oral Reading Fluency	75.40	40.40	0	45	71	102	219	2133

Note. 25th = 1st quartile; 50th = 2nd quartile; 75th = 3rd quartile.

February 18, 2008

DIBELS Summit, Albuquerque, NM

55

NWF-Middle of 1st Grade

Table X

Likelihood of Achieving Benchmark Outcomes for Decisions Based on NWF First Grade Middle of Year Scores

	ORF MOY	NWF EOY	ORF EOY
Likelihood of achieving benchmark outcomes			
Low Risk: NWF score is 50 or more	.88	.79	.96
Some Risk: NWF score is 30 to 49	.48	.35	.71
High Risk: NWF score is 0 to 29	.14	.07	.27
Area under the ROC curve			
Low risk score on outcome	.86	.84	.87
High risk score on outcome	.93	.86	.92

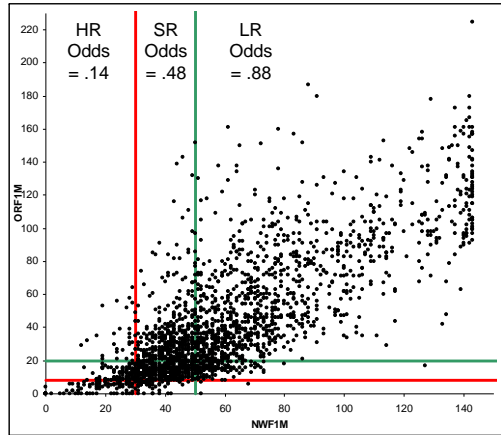
Note. Likelihood is reported as a conditional probability of a low risk outcome given NWF MOY score. NWF = Nonsense Word Fluency; ORF = Oral Reading Fluency; MOY = Middle of Year; EOY = End of Year; ROC = Receiver Operator Characteristic.

February 18, 2008

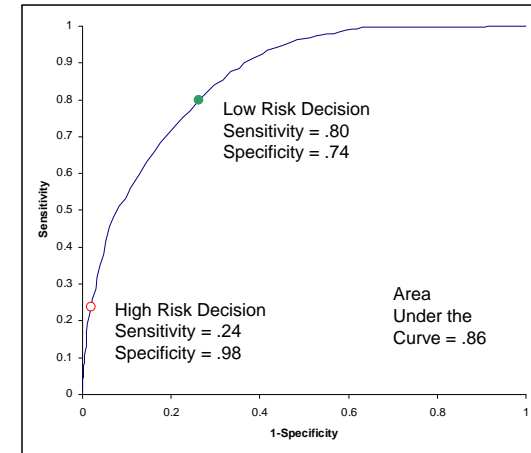
DIBELS Summit, Albuquerque, NM

56

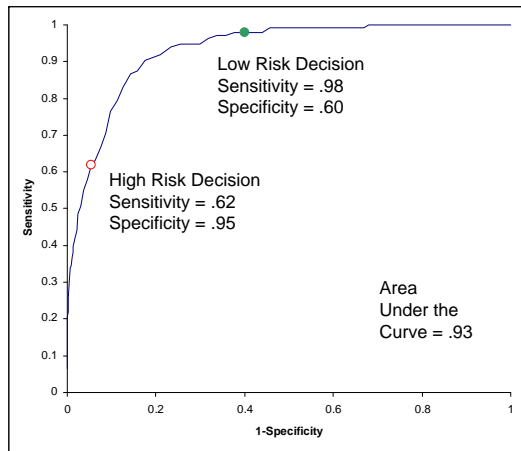
NWF1M to ORF1E



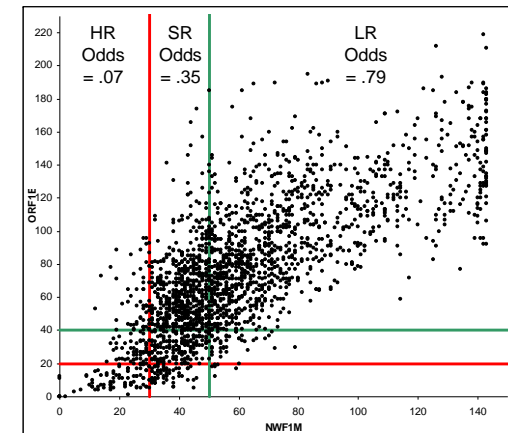
ROC for First Grade, Middle of Year ORF Low Risk Outcome



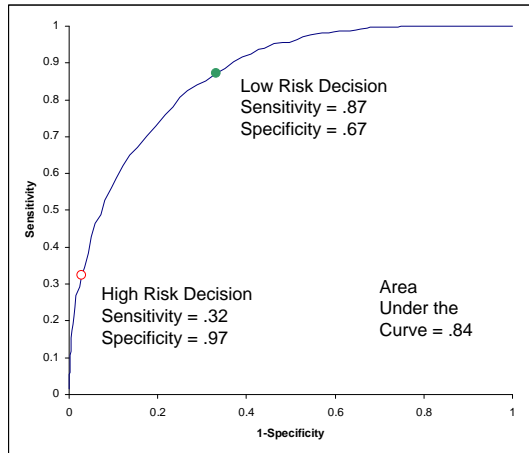
ROC for First Grade, Middle of Year ORF High Risk Outcome



NWF1M to ORF1E



ROC for First Grade, End of Year ORF Low Risk Outcome

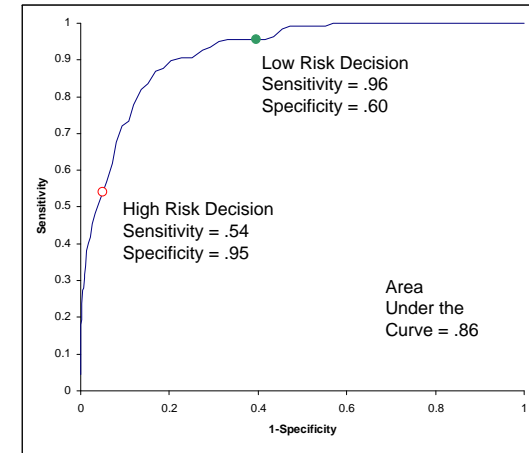


February 18, 2008

DIBELS Summit, Albuquerque, NM

61

ROC for First Grade, End of Year ORF High Risk Outcome

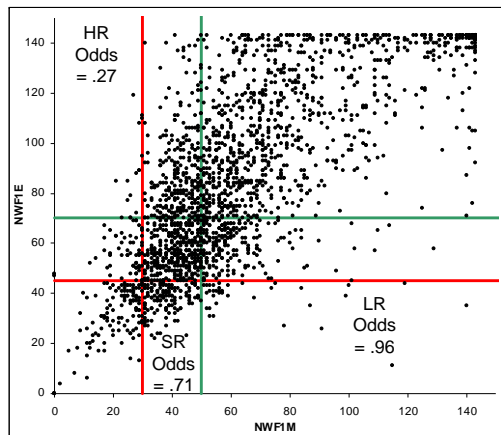


February 18, 2008

DIBELS Summit, Albuquerque, NM

62

NWF1M to NWF1E

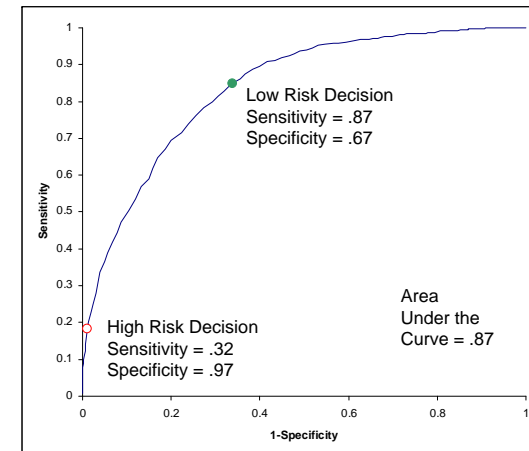


February 18, 2008

DIBELS Summit, Albuquerque, NM

63

ROC for First Grade, End of Year NWF Low Risk Outcome

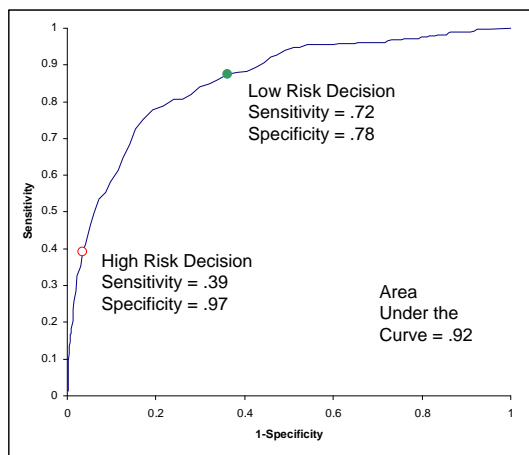


February 18, 2008

DIBELS Summit, Albuquerque, NM

64

ROC for First Grade, End of Year NWF High Risk Outcome



February 18, 2008

DIBELS Summit, Albuquerque, NM

65

NWF-End of 1st Grade

Descriptive Statistics for DIBELS First Grade Measures

Measure	Mean	SD	Min	25th	50th	75th	Max	N
End of year								
Nonsense Word Fluency	81.83	34.45	0	55	76	108	143	2135
Oral Reading Fluency	75.40	40.40	0	45	71	102	219	2133

Note. 25th = 1st quartile; 50th = 2nd quartile; 75th = 3rd quartile. Correlation between Nonsense Word Fluency end of year and Oral Reading Fluency end of year scores is .77(2133), $p < .01$; the number of subjects with pairwise complete data is reported in parentheses

February 18, 2008

DIBELS Summit, Albuquerque, NM

66

NWF-End of 1st Grade

Likelihood of Achieving ORF First Grade End of Year Benchmark Outcomes for Decisions Based on NWF First Grade End of Year Scores

Likelihood of achieving benchmark outcomes	
Low Risk: NWF score is 70 or more	.97
Some Risk: NWF score is 44 to 69	.72
High Risk: NWF score is 0 to 43	.32
Area under the ROC curve	
Low risk score on outcome	.90
High risk score on outcome	.93

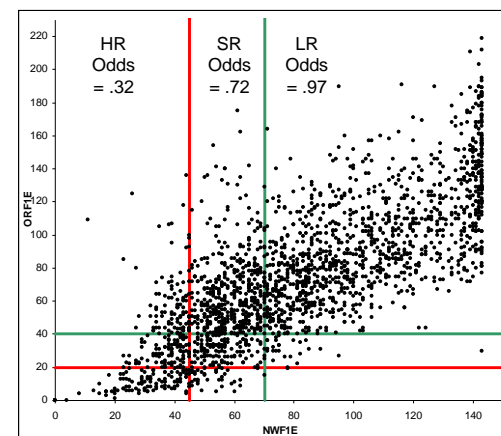
Note. Likelihood is reported as a conditional probability of a low risk outcome given NWF EOY score. NWF = Nonsense Word Fluency; ORF = Oral Reading Fluency; BOY = Beginning of Year; MOY = Middle of Year; EOY = End of Year; ROC = Receiver Operator Characteristic.

February 18, 2008

DIBELS Summit, Albuquerque, NM

67

NWF1E to ORF1E

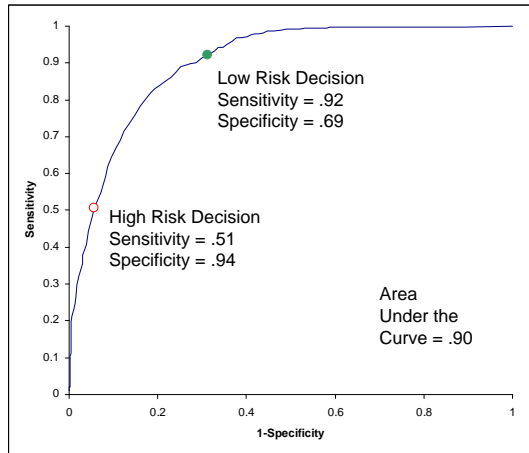


February 18, 2008

DIBELS Summit, Albuquerque, NM

68

ROC for First Grade, End of Year ORF Low Risk Outcome

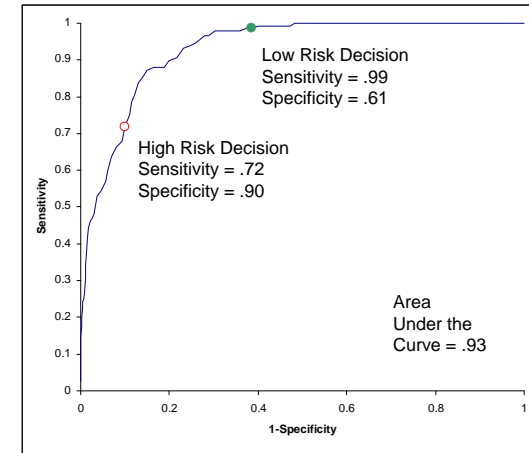


February 18, 2008

DIBELS Summit, Albuquerque, NM

69

ROC for First Grade, End of Year ORF High Risk Outcome



February 18, 2008

DIBELS Summit, Albuquerque, NM

70

NWF-Beginning of 2nd Grade

Table X

Descriptive Statistics for DIBELS Second Grade Measures

Measure	Mean	SD	Min	25th	50th	75th	Max	N
Beginning of year								
Nonsense Word Fluency	79.43	36.85	0	49	74	110	143	2254
Oral Reading Fluency	58.22	35.59	0	31	51	79	201	1096
Middle of year								
Oral Reading Fluency	83.68	38.17	0	57	81	109	209	1098
End of year								
Oral Reading Fluency	107.57	39.65	0	80	108	133	251	2257

Note. 25th = 1st quartile; 50th = 2nd quartile; 75th = 3rd quartile.

February 18, 2008

DIBELS Summit, Albuquerque, NM

71

NWF-Beginning of 2nd Grade

Table X

Likelihood of Achieving Benchmark Outcomes for Decisions Based on NWF Second Grade Beginning of Year Scores

	ORF BOY	ORF MOY	ORF EOY
Likelihood of achieving benchmark outcomes			
Low Risk: NWF score is 70 or more	.94	.95	.92
Some Risk: NWF score is 45 to 69	.51	.58	.47
High Risk: NWF score is 0 to 44	.21	.30	.22
Area under the ROC curve			
Low risk score on outcome	0.90	0.89	0.88
High risk score on outcome	0.91	0.89	0.89

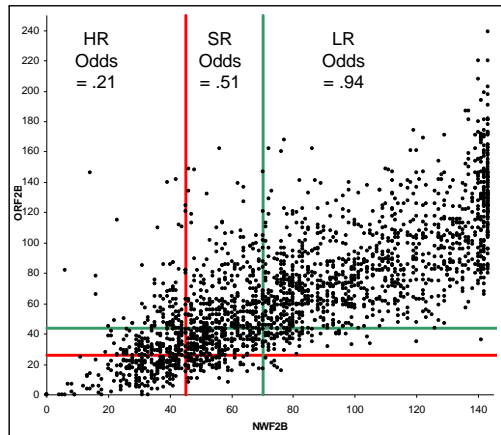
Note. Likelihood is reported as a conditional probability of a low risk outcome given NWF BOY score. NWF = Nonsense Word Fluency; ORF = Oral Reading Fluency; BOY = Beginning of Year; MOY = Middle of Year; EOY = End of Year; ROC = Receiver Operator Characteristic.

February 18, 2008

DIBELS Summit, Albuquerque, NM

72

NWF2B to ORF2B

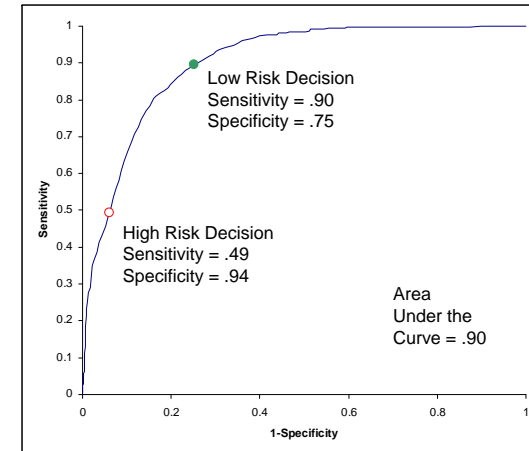


February 18, 2008

DIBELS Summit, Albuquerque, NM

73

ROC for Second Grade, Beginning of Year ORF Low Risk Outcome

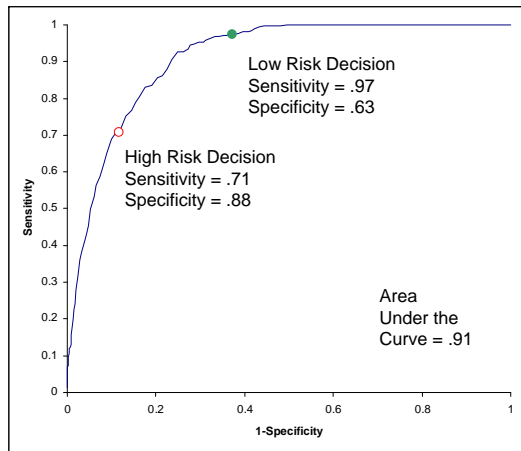


February 18, 2008

DIBELS Summit, Albuquerque, NM

74

ROC for Second Grade, Beginning of Year ORF High Risk Outcome

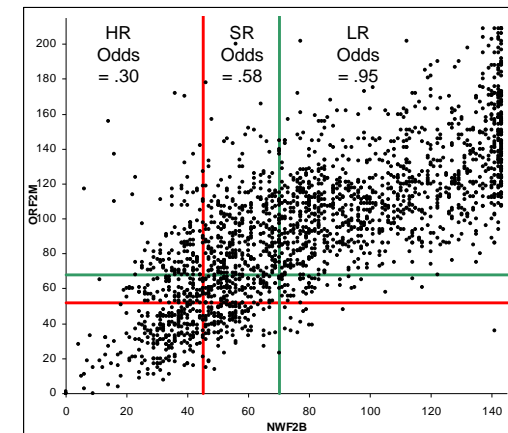


February 18, 2008

DIBELS Summit, Albuquerque, NM

75

NWF2B to ORF2M

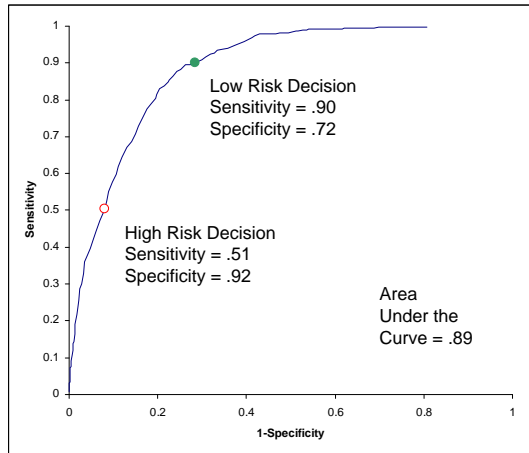


February 18, 2008

DIBELS Summit, Albuquerque, NM

76

ROC for Second Grade, Middle of Year ORF Low Risk Outcome

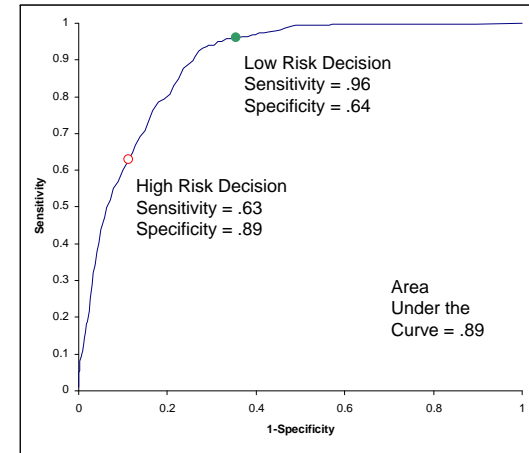


February 18, 2008

DIBELS Summit, Albuquerque, NM

77

ROC for Second Grade, Middle of Year ORF High Risk Outcome

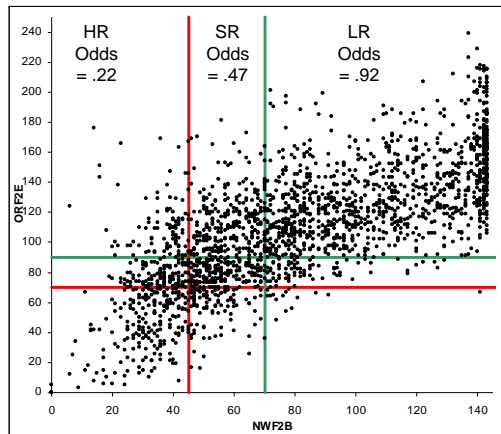


February 18, 2008

DIBELS Summit, Albuquerque, NM

78

NWF2B to ORF2E

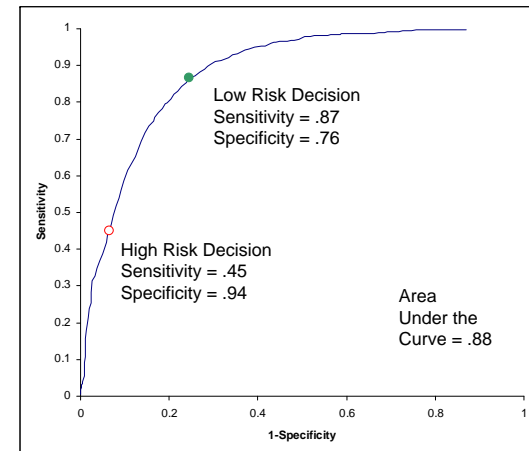


February 18, 2008

DIBELS Summit, Albuquerque, NM

79

ROC for Second Grade, End of Year ORF Low Risk Outcome

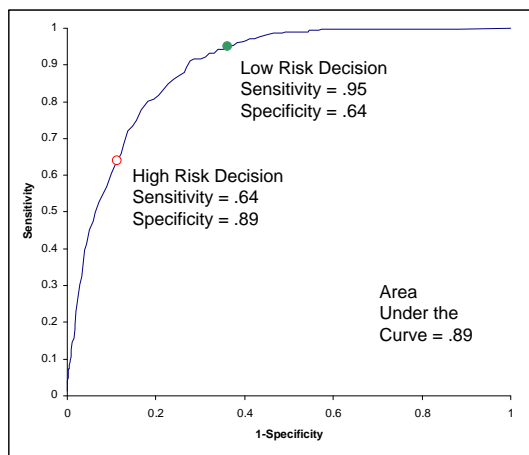


February 18, 2008

DIBELS Summit, Albuquerque, NM

80

ROC for Second Grade, End of Year ORF High Risk Outcome



ORF ROC Curves, for Second Grade

ORF-Beginning of 2nd Grade

Table X

Descriptive Statistics for DIBELS Second Grade Measures

Measure	Mean	SD	Min	25th	50th	75th	Max	N
Beginning of year								
Oral Reading Fluency	58.22	35.59	0	31	51	79	201	1096
Middle of year								
Oral Reading Fluency	83.68	38.17	0	57	81	109	209	1098
End of year								
Oral Reading Fluency	107.57	39.65	0	80	108	133	251	2257

Note. 25th = 1st quartile; 50th = 2nd quartile; 75th = 3rd quartile.

ORF-Beginning of 2nd Grade

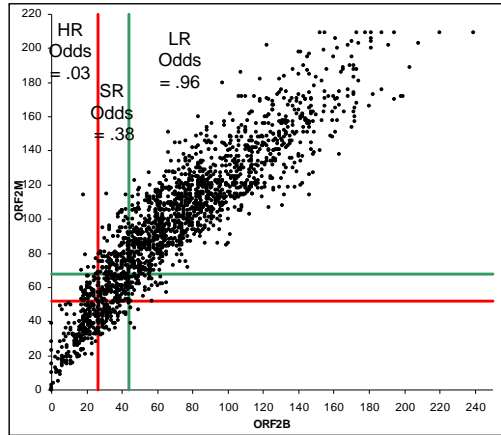
Table X

Likelihood of Achieving Benchmark Outcomes for Decisions Based on ORF Second Grade Beginning of Year Scores

	ORF MOY	ORF EOY
Likelihood of achieving benchmark outcomes		
Low Risk: ORF score is 44 or more	.96	.91
Some Risk: ORF score is 26 to 43	.38	.23
High Risk: ORF score is 0 to 25	.03	.03
Area under the ROC curve		
Low risk score on outcome	.97	.96
High risk score on outcome	.97	.96

Note. Likelihood is reported as a conditional probability of a low risk outcome given ORF BOY score. ORF = Oral Reading Fluency; BOY = Beginning of Year; MOY = Middle of Year; EOY = End of Year; ROC = Receiver Operator Characteristic.

ORF2B to ORF2M

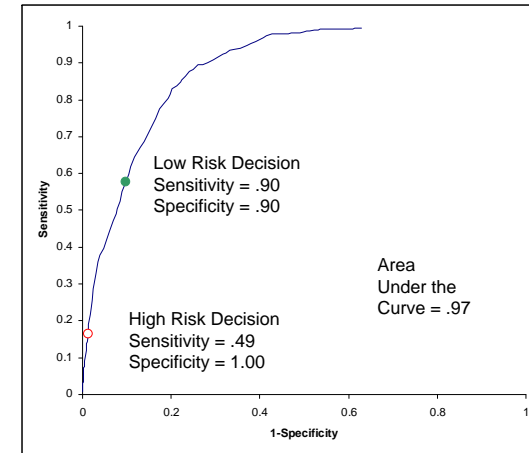


February 18, 2008

DIBELS Summit, Albuquerque, NM

85

ROC for Second Grade, Middle of Year ORF Low Risk Outcome

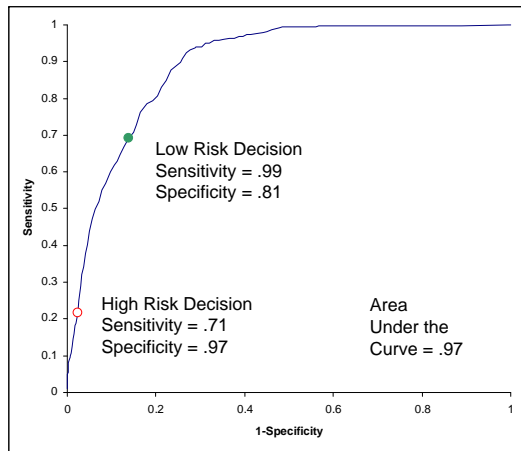


February 18, 2008

DIBELS Summit, Albuquerque, NM

86

ROC for Second Grade, Middle of Year ORF High Risk Outcome

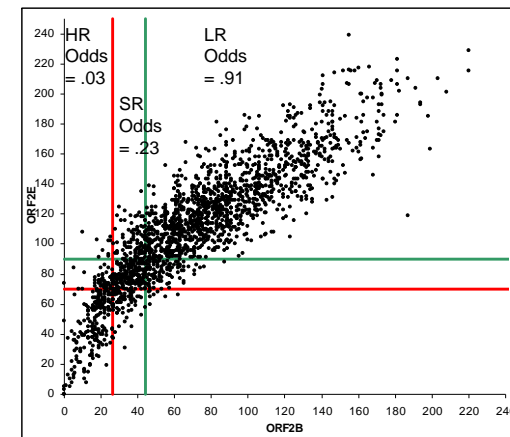


February 18, 2008

DIBELS Summit, Albuquerque, NM

87

ORF2B to ORF2E

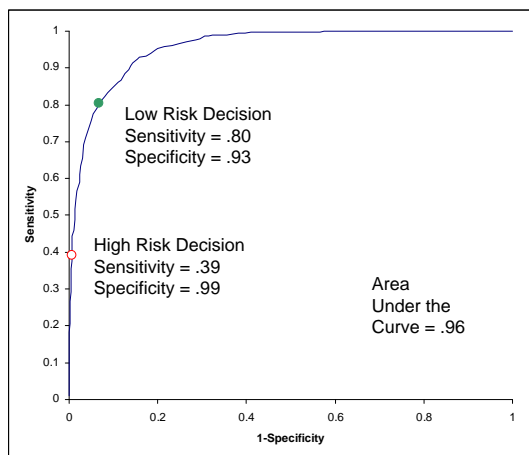


February 18, 2008

DIBELS Summit, Albuquerque, NM

88

ROC for Second Grade, End of Year ORF Low Risk Outcome

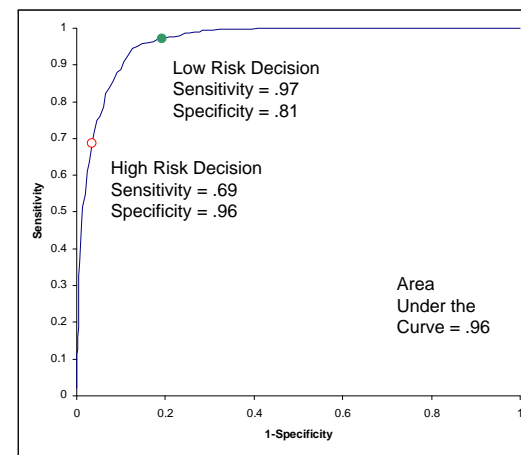


February 18, 2008

DIBELS Summit, Albuquerque, NM

89

ROC for Second Grade, End of Year ORF High Risk Outcome



February 18, 2008

DIBELS Summit, Albuquerque, NM

90

ORF-Middle of 2nd Grade

Table X

Descriptive Statistics for DIBELS Second Grade Measures

Measure	Mean	SD	Min	25th	50th	75th	Max	N
Middle of year								
Oral Reading Fluency	83.68	38.17	0	57	81	109	209	1098
End of year								
Oral Reading Fluency	107.57	39.65	0	80	108	133	251	2257

Note. 25th = 1st quartile; 50th = 2nd quartile; 75th = 3rd quartile. Correlation between Oral Reading Fluency middle and end of year scores is .92(1070), $p < .01$; the number of subjects with pair-wise complete data is reported in parentheses.

February 18, 2008

DIBELS Summit, Albuquerque, NM

91

ORF-Middle of 2nd Grade

Table X

Likelihood of Achieving Benchmark Outcomes for Decisions Based on ORF Second Grade Middle of Year Scores

	ORF EOY
Likelihood of achieving benchmark outcomes	
Low Risk: ORF score is 68 or more	.89
Some Risk: ORF score is 52 to 67	.19
High Risk: ORF score is 0 to 51	.01
Area under the ROC curve	
Low risk score on outcome	.97
High risk score on outcome	.98

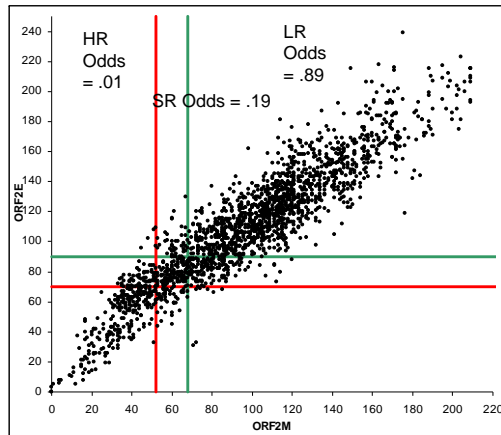
Note. Likelihood is reported as a conditional probability of a low risk outcome given ORF MOY score. ORF = Oral Reading Fluency; MOY = Middle of Year; EOY = End of Year; ROC = Receiver Operator Characteristic.

February 18, 2008

DIBELS Summit, Albuquerque, NM

92

ORF2M to ORF2E

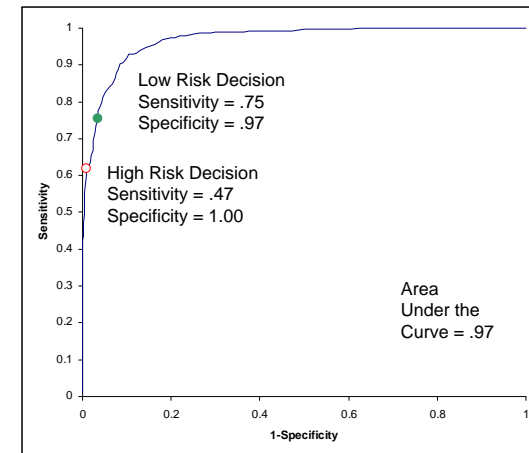


February 18, 2008

DIBELS Summit, Albuquerque, NM

93

ROC for Second Grade, End of Year ORF Low Risk Outcome

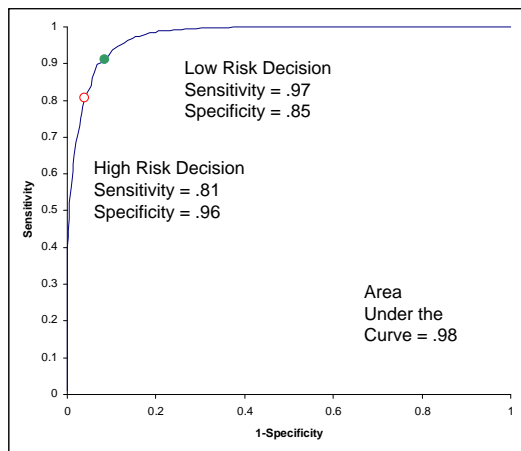


February 18, 2008

DIBELS Summit, Albuquerque, NM

94

ROC for Second Grade, End of Year ORF High Risk Outcome



February 18, 2008

DIBELS Summit, Albuquerque, NM

95

Findings from Beta 1 Study

- Validity correlation coefficients are strong, and positive, with clear patterns emerging by construct. NWF and ORF correlated slightly higher with each other than with measures of phonemic awareness (i.e. FSF, PSF).
- Benchmark goals and cutpoints function according to the design specifications.
- AUC coefficients were exceptional (.79 – .98; over half were above .90!), across all predictors, even when more distal outcomes were used.

February 18, 2008

DIBELS Summit, Albuquerque, NM

96

Using DIBELS Benchmark Goals and Cutpoints: Recommendations

- For all measures, the primary goal is meaningful. Delivering effective, appropriate, differentiated instruction that is cohesive and integrated is the key to reaching this marker for your students.
- However, the powerful predictive validity of the measures does not mean that they should become proxies for other, high stakes, assessments.

Using DIBELS Benchmark Goals and Cutpoints: Recommendations

- DIBELS Benchmark goals and cutpoints can represent meaningful and important goals for progress monitoring.
 - These goals are based on a national norm
 - These goals are referenced to both “internal” criteria (Oral Reading Fluency) and “external” criteria (state tests)
- The goals can also be used to evaluate your overall system of support.
 - We should spend as much time evaluating our instruction as we do child’s response to it.

Questions?

- Kelli D. Cummings, Ph.D., NCSP
kcummings@dibels.org
- Roland H. Good III, Ph.D.
rhgood@dibels.org