ROC Done Right! Using ROC Curve Analyses to Enhance Prevention

Kelli D. Cummings, Ph.D., NCSP
Kelly A. Powell-Smith, Ph.D., NCSP

NASP
February 27th, 2009

Dynamic Measurement Group
Supporting School Success One Step at a Time

Acknowledgements

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

Disclosure Statement

- Kelli Cummings & Kelly Powell-Smith are employees of Dynamic Measurement Group, the home of DIBELS® research & development, and may benefit financially from the sales of DIBELS and DIBELS-related products and services.

Overview

- Introduction
- Evaluating screening tools in an instructional context is critical.
- Two common metrics, sensitivity and specificity, are problematic in this context.
- Conditional Probabilities may offer a way to assess a measure's validity and are interpretable.
  - Example using DIBELS® NWF
- Discussion
- Questions and Answers

Key Terms

- Area Under the Curve
- True Positive
- False Positive
- True Negative
- False Negative

- Sensitivity
- Specificity
- Positive Predictive Power
- Negative Predictive Power
- Classification Accuracy

Questions To Ponder

- Where are we in our state of knowledge?
- Where are we in our current technology?
- What are we trying to accomplish?
- What level of error are we willing to tolerate?

Relevance of ROC Curve Analysis School Psychologists

- Educational Milieu & Changes in Practice
  - Accountability
  - Response to Intervention (RTI)
  - Need for instruments useful for screening, goal setting, and progress monitoring
- Increased attention to diagnostic utility and specifically the use of ROC

What School Psychologists Need To Know To Be Good Consumers?

- The purpose of screening
- The impact of context
  - Relevant concerns with evaluating diagnostic utility within a preventative educational context
- Different diagnostic utility metrics
- How ROC curves fit in the picture

The Purpose of Screening Tools in Education

- Quickly identify the likelihood a student will need additional help to *prevent* a later academic difficulty.
- Specify important & meaningful goals—a point at which we change the odds to being in favor of an individual's meeting subsequent goals.

The Impact of Context

- Key Point: Outcomes are unknown and are likely ***not even present*** at the time of the screening.

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.

Critical Evaluation of Screening Tools

- We need to evaluate the:
  - Reliability & validity of the measures,
  - Decision utility of the measures,
  - 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.

Screening for Tuberculosis

	Screening Decision: Positive TB	Negative TB
True State (Outcome): Negative for tuberculosis	FP: False Positive	TN: True Negative
True State (Outcome): Positive for tuberculosis	TP: True Positive	FN: False Negative

- ***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

- Our recommendation is to use the likelihood of achieving important educational outcomes because:
  - The outcome is continuous.
  - There is a lack of general agreement on a specific assessment or cutpoint on an assessment that discriminates adequate and not adequate skills.
  - And especially because of intervening instruction and occurring between the screening assessment and the outcome. **When there is 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? $\dfrac{n_{13}}{n_{13} + n_{23} + n_{33}}$
- ***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? $\dfrac{n_{12}}{n_{12} + n_{22} + n_{32}}$
- ***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? $\dfrac{n_{11}}{n_{11} + n_{21} + n_{31}}$

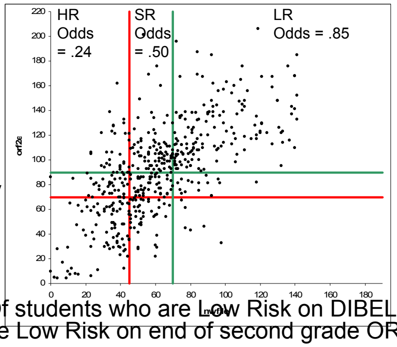
For Example, DIBELS Assessment

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

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

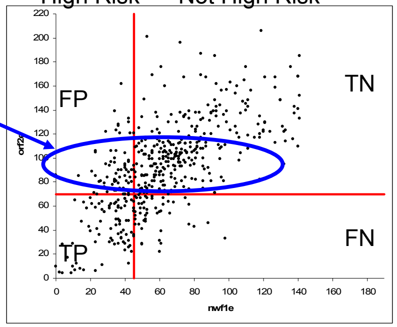
We can impose a 2-by-2 Model on Reading Assessment, but it Doesn't Really Fit

What about these students?

DIBELS Alphabetic Principle:
High Risk Not High Risk

Second End ORF Outcome:
Not High Risk

High Risk

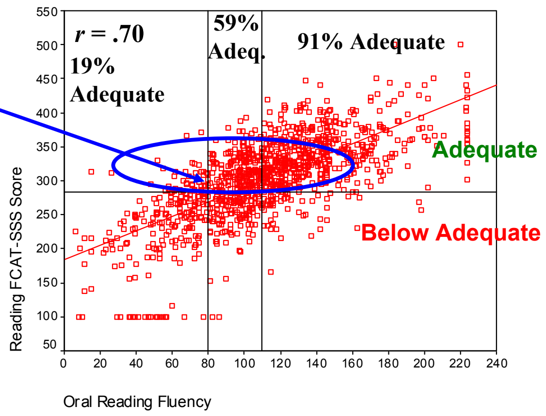


- ***Sensitivity***: Of students ***who truly have poor reading***, what proportion are identified as having poor reading by DIBELS? $\frac{TP}{TP + FN}$
- ***Specificity***: Of students ***who truly do not have poor reading***, what proportion are identified as not having poor reading on DIBELS? $\frac{TN}{FP + TN}$

Any Two, High Quality Reading Criterion Tests Have a Zone of Disagreement

What about these students?

Between G3 ORF of 80 and 110, the odds are 59% the student will rank "adequate" on the FL State Assessment.



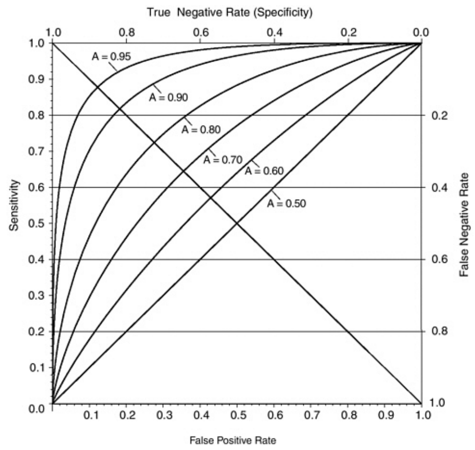
Buck, J., & Torgesen, J. (2003). *The relationship between performance on a measure of oral reading fluency and performance on the Florida Comprehensive Assessment Test* (Technical Report 1). Tallahassee, FL: Florida Center for Reading Research,.

How do we define at-risk reading outcomes?

Study	Outcome Criterion	Outcome Test	Time of Year
Foorman et al. (1998)	<23rd Percentile	WJ-R Broad Reading	Spring of 1st
""	*Not specified*	WJ-R Broad Reading	Spring of 1st
""	<36th Percentile	WJ Broad Reading	Spring of 2nd
O'Connor & Jenkins (1999)	<8th Percentile	WRMT BRS	1st
Speece et al. (2003)	<26th Percentile	WJ-R Word Attack	Spring of 1st
""	<26th Percentile	CBM ORF	Spring of 1st
Schatschneider (2006)	<25th Percentile	SAT-10 RC	Spring of 1st
""	<25th Percentile	SAT-10 RC	Spring of 2nd
""	<Level 3	FCAT RC	Spring of 3rd
Good et al. (2001)	<40 WRC	CBM ORF	Spring of 1st
""	<50 WRC	CBM ORF	Spring of 2nd
""	"Does not meet expectations"	OSA	Spring of 3rd
Speece & Case (2001)	DD (-1 SD on slope & level)	CBM ORF	*Not specified*
Speece (2005)	<40 WRC & -1 SD slope	CBM ORF	Spring of 1st
Compton et al. (2006)	<85 SS	Broad Reading Composite	Spring of 2nd
""	<85 SS	Component Reading	Spring of 2nd
Good et al. (in-press)	<40 WRC	DIBELS ORF	Spring of 1st
Stage & Jacobsen (2001)	"Below proficiency"	WASL RC	*Not specified*
McGlinchey & Hixson (2004)	"Below proficiency"	MEAP	*Not specified*

Note. This table adapted from Jenkins, Hudson, & Johnson (2007). Screening for at-risk readers in a response to intervention framework *School Psychology Review*, *36*(4), 582-600.

Increasing Focus on ROC Curves



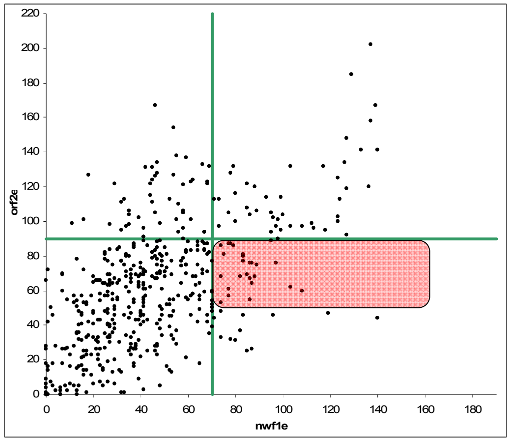
Some Problems Along the Way

- Development of cut-scores
  - in the absence of distinguishing between the metrics used to evaluate them
  - in the absence of understanding the implications of their use (consequences)
  - because doing so is fashionable
- Evaluation of screening tools and associated cut scores in the absence of discussing the role of the context in which the tools are used.
- Failure to acknowledge the lack of agreement on a gold standard.

What are the Implications of Statements About High "False Positive" Rates?

- Are we providing intervention to students who don't really need it?
- OR
  - Did those students receive some very high quality instruction or intensive intervention?
  - Is the outcome measure not very good (e.g., poor technical properties or too easy)?
  - Did the child have a bad minute?

Sensitivity & Specificity Logic Doesn't Work



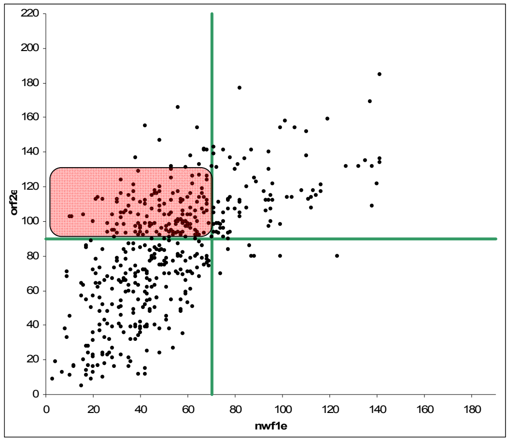
Sample District 1

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.
- Do we really want to consider these students to be "False Negatives"? Or are they failures of our Tier 1 instruction?

Note. Outcome baserate would be .80.

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?

Note. Outcome baserate would be .49.

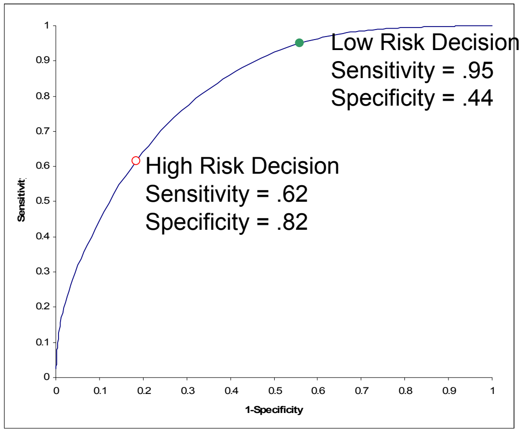
How Do You Decide Which Explanation Is Correct?

- Did the student get good instruction?
  - Document instructional context
- Are the outcome measures adequate?
  - Critically evaluate the outcome measures used
- Did the child have a bad minute?
  - Validate the score by retesting

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.

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



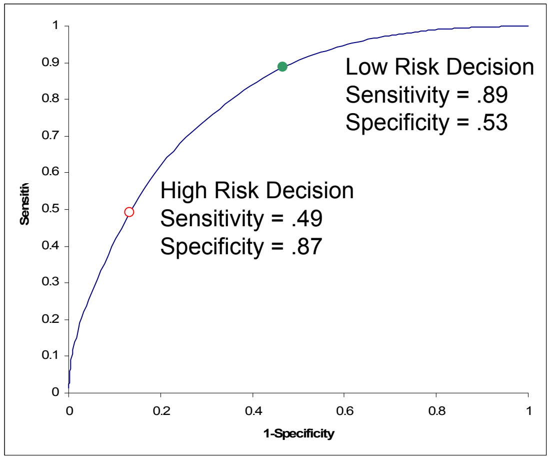
Low Risk Decision
Sensitivity = .95
Specificity = .44

High Risk Decision
Sensitivity = .62
Specificity = .82

Area
Under the
Curve = .80

Full WG Sample, *n* = 58811

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



Low Risk Decision
Sensitivity = .89
Specificity = .53

High Risk Decision
Sensitivity = .49
Specificity = .87

Area
Under the
Curve = .82

Full WG Sample, *n* = 58811

Summary

- Lack of agreement on the outcome test
- Lack of agreement on the outcome level
- There is a zone of uncertainty
- Consider the impact of instructional context

Our recommendations: Use likelihood of achieving important educational outcomes. Choose an outcome measure that is meaningful and important. Compare AUCs across potential screeners. Examine national sample to determine what it takes to put the odds in a student's favor.

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.

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.

Evaluating Screening Measures in Psychology and Education What do you need to know?

- Reliability
- Validity
  - Concurrent validity
  - Predictive validity
  - Treatment utility
  - Social validity

Evaluating a Tool? Start Here!

- Use SAS PROC Logistic to generate a ROC Curve
- Examine the area under the curve (AUC) to assess the measures’ content validity
- Cut and paste raw output into Excel to draw a ROC Curve
- Use raw data and cut scores to evaluate conditional probabilities (“odds”)

SAS syntax: Using cutpoints to assign risk values and create temporary data sets.

```
/*  Status of 0 means "On Track" or "Low Risk"
      g1stat = Low Risk / Not Low Risk (Some + At Risk)
      g1rstat = At Risk / Not At Risk (Some + Low Risk)        */
–    data g133 g133sub;
–      set saslib.beta;
–      if orf1e ge 40 then g1stat=0; if orf1e lt 40 then g1stat=1;
–      if orf1e lt 20 then g1rstat=1; if orf1e ge 20 then g1rstat=0;
–      output g133;
/* For all status categories on individual measures, 3 = Benchmark / Low Risk;
    2 = Strategic / Some Risk; 1 = Intensive / At Risk            */

–      if nwf1e ge 70 then nwfesta=3;
–      if nwf1e ge 45 and nwf1e lt 70 then nwfesta=2;
–      if nwf1e lt 45 then nwfesta=1;

–      if orf1e ge 40 then orfesta=3;
–      if orf1e ge 20 and orf1e lt 40 then orfesta=2;
–      if orf1e lt 20 then orfesta=1;

–      if n(nwf1e, orf1e)=2 then output g133sub;
–    run;
```

SAS syntax: Scatterplot and contingency table.

```
/*This sample data is what gets pasted in to Excel to create the scatterplots. Make sure
    that data are always output in the form of
      student(bid)           predictor value                     outcome value       */
–   data nwf1eorf1e;
–     set g133sub;
–     file outdat;
–     put bid nwf1e orf1e ;
/*Examines the contingency table (3X3)
      1 = Intensive
      2 = Strategic
      3 = Benchmark                    */

–   proc freq data=g133sub;
/*This syntax is used to determine the effect of instructional context on outcome status,
    e.g. how many at-risk students per district at EOY.*/
–   proc corr data=g133sub;
–     var  nwf1e orf1e;
–     run;
```

SAS syntax: Creating the Low Risk and High Risk ROC Curves.

```
/*Computes the regression line for NWF1e predicting ORF1e outcomes.*/
proc glm data=g133sub;
  model orf1e = nwf1e  / solution;
  run;
/*Runs the ROC Curve for NWF1e predicting ORF1e "on-track" status b = "benchmark status"*/
proc logistic data=g133sub;
  model g1stat(event='1') = nwf1e / outroc=roc1b ;
  title 'Low Risk ROC';
run;
/*Runs the ROC Curve for NWF1e predicting ORF1e "at-risk" status r = "risk status"*/
proc logistic data=g133sub;
  model g1rstat(event='1') = nwf1e / outroc=roc1r ;
  title 'High Risk ROC';
run;
proc print data=roc1b;
title 'Low Risk ROC';
proc print data=roc1r;
title 'High Risk ROC';
run;
```

SAS output: Sample for scatterplot.

As specified from the SAS syntax, this sample data is in the form of
student(bid) predictor value(NWF1E) outcome value(ORF1E)
and gets pasted into Excel to create the scatterplots.

- 1010000011858 85 71
- 1010000011948 42 38
- 1010000011949 109 56
- 1010000011950 65 30
- 1010000011951 88 42
- 1010000011953 143 154
- 1010000011954 31 8
- 1010000011955 62 42
- 1010000011957 60 40
- 1010000011958 83 42
- .

SAS output: Low Risk ROC Curve

Obs	Probability	True	True	False	False	Sensitivity	1-Specificity
		Positive	Negative	Positive	Negative		
1	0.97750	4	1726	0	403	0.00983	0.00000
2	0.97127	5	1726	0	402	0.01229	0.00000
3	0.96884	6	1726	0	401	0.01474	0.00000
4	0.95700	7	1726	0	400	0.01720	0.00000
5	0.94541	7	1725	1	400	0.01720	0.00058
6	0.93611	8	1725	1	399	0.01966	0.00058
7	0.93093	8	1724	2	399	0.01966	0.00116
8	0.92535	10	1724	2	397	0.02457	0.00116
9	0.91295	13	1724	2	394	0.03194	0.00116
10	0.90608	14	1724	2	393	0.03440	0.00116

SAS output: High Risk ROC Curve

Obs	Probability	True	True	False	False	Sensitivity	1-Specificity
		Positive	Negative	Positive	Negative		
1	0.96586	4	1984	0	145	0.02685	0.00000
2	0.95286	5	1984	0	144	0.03356	0.00000
3	0.94756	6	1984	0	143	0.04027	0.00000
4	0.92027	7	1984	0	142	0.04698	0.00000
5	0.89185	7	1983	1	142	0.04698	0.00050
6	0.86827	8	1983	1	141	0.05369	0.00050
7	0.85491	8	1982	2	141	0.05369	0.00101
8	0.84045	10	1982	2	139	0.06711	0.00101
9	0.80807	13	1982	2	136	0.08725	0.00101
10	0.79009	14	1982	2	135	0.09396	0.00101

SAS output: Proc Logistic

```
Low Risk ROC:
Association of Predicted Probabilities and Observed Responses
Percent Concordant     89.6   Somers' D    0.800
Percent Discordant      9.7   Gamma        0.806
Percent Tied            0.7   Tau-a        0.247
Pairs                702482   c            0.900

High Risk ROC:
Association of Predicted Probabilities and Observed Responses
Percent Concordant     92.7   Somers' D    0.859
Percent Discordant      6.7   Gamma        0.864
Percent Tied            0.6   Tau-a        0.112
Pairs                295616   c            0.930
```

Evaluating a Tool?

- Use raw data and cut scores to evaluate sensitivity/specificity, PPP/NPP but remember ..
- You will have to apply a two-by-two logic to a three-by-three world.

So for example

- 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
n	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
sd	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
sd	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.

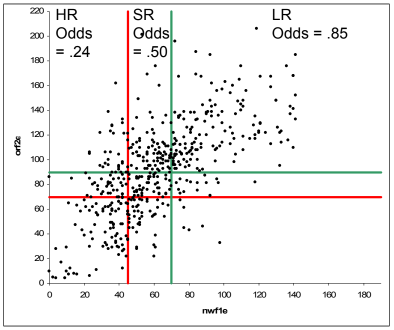
High Risk, Some Risk, and Low Risk Decisions

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



- 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

When evaluating DIBELS

- We ask that you consider the primary design specifications and compare your results accordingly.
- We ask that you consider all diagnostic utility stats, paying primary attention to AUC as an evaluation of the *measure*.
- All other diagnostic utility stats depend on cutpoint, so we ask that you note how and why the scores were force-dichotomized as you analyzed the data.

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-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 pair-wise complete data is reported in parentheses

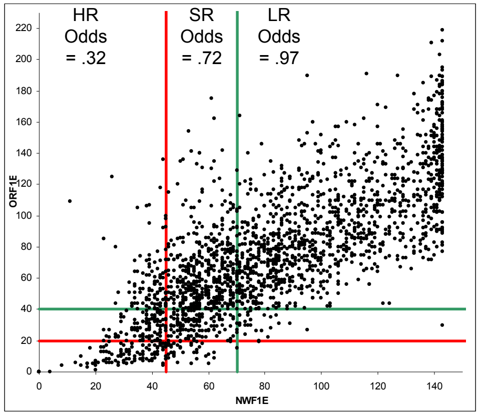
NWF-End of 1st Grade

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

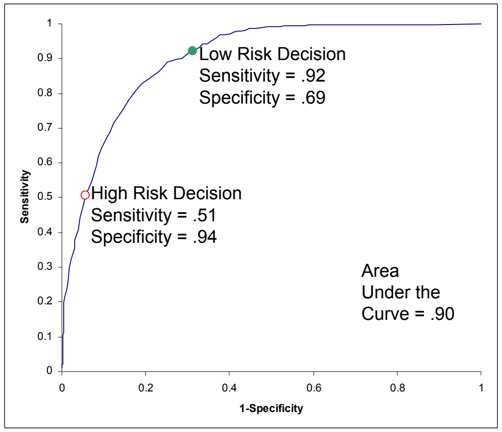
Likelihood of achieving benchmark outcomes	
Low Risk: NWF score is 70 or more	.97
Some Risk: NWF score is 44 to 69	.72
High Risk: NWF score is 0 to 45	.32
Area under the ROC curve	
Low risk score on outcome	.90
High risk score on outcome	.93

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.

NWF1E to ORF1E



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

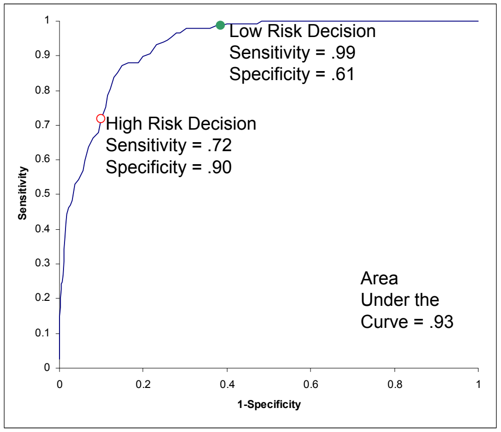


Low Risk Decision
Sensitivity = .92
Specificity = .69

High Risk Decision
Sensitivity = .51
Specificity = .94

Area
Under the
Curve = .90

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



Low Risk Decision
Sensitivity = .99
Specificity = .61

High Risk Decision
Sensitivity = .72
Specificity = .90

Area
Under the
Curve = .93

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.

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.

ROC Done Right?

Using a Tool for Screening or Progress Monitoring? Consider the following:

- Is the tool reliable? (Same standards apply, Salvia & Ysseldyke, 2009)
- Is the tool valid?
  - Are there high concurrent correlations?
  - Are there high predictive correlations? (To skills that are distal? To similar skills that are measured at distal time points?)
  - Does the tool add value as a predictor? (AUC)

Using a Tool for Screening or Progress Monitoring? Consider the following:

- Is the tool valid?
  - Treatment validity. Do the scores represent meaningful goals that change outcomes? Do the scores help you to make decisions about individuals? Systems?
  - Social validity. Does improvement on the measures, and attainment of the goal, make a difference to individual students? To their teachers? To their parents?

We must critically evaluate our screening tools

- However
- 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.

Questions?

- Kelli D. Cummings, Ph.D., NCSP

  kcummings@dibels.org

- Kelly A. Powell-Smith, Ph.D., NCSP

  kpowellsmith@dibels.org