ROC Done Right? Part 3
DIBELS® Benchmark Goals

Pacific Coast Research Conference
Coronado, CA
February 9, 2008

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

Kelli D. Cummings
Kelly A. Powell-Smith
Dynamic Measurement Group, Inc.

Part 3 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.
- Additional problems of Sensitivity and Specificity:
  - They depend on the choice of cutpoint. (ROC curves address the problem of different cutpoints, but ROC curves don't address the other problems of sensitivity and specificity).
  - They are affected by differences in baserate.
  - They are affected by differences in Tier 1 instruction and Tier 2 and Tier 3 interventions.
- Our recommendation:
  - Use likelihood or odds of achieving important educational outcomes to evaluate screening assessments.

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, with 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.

Part 3 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.
- Additional problems of Sensitivity and Specificity:
  - They depend on the choice of cutpoint. (ROC curves address the problem of different cutpoints, but ROC curves don't address the other problems of sensitivity and specificity).
  - They are affected by differences in baserate.
  - They are affected by differences in Tier 1 instruction and Tier 2 and Tier 3 interventions.
- 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,
  - 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.

For Example, Screening for Tuberculosis

	Screening Decision: Positive TB	Screening Decision: 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? $\frac{TP}{TP + FN}$
- ***Specificity***: Of individuals ***who truly do not have tuberculosis***, what proportion are identified as not having tuberculosis on the screening test? $\frac{TN}{FP + TN}$

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.

Part 3 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.
- Additional problems of Sensitivity and Specificity:
  - They depend on the choice of cutpoint. (ROC curves address the problem of different cutpoints, but ROC curves don't address the other problems of sensitivity and specificity).
  - They are affected by differences in baserate.
  - They are affected by differences in Tier 1 instruction and Tier 2 and Tier 3 interventions.
- Our recommendation:
  - Use likelihood or odds of achieving important educational outcomes to evaluate screening assessments.

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? $\frac{n_{11}}{n_{11}+n_{21}+n_{31}}$

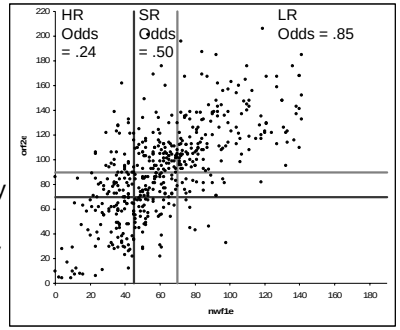
For Example, DIBELS Assessment

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

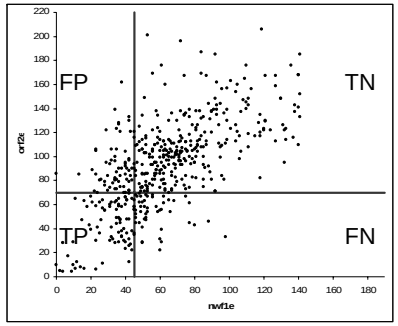
Note: Odds based on Full WG sample, *n* = 58811. Scatterplot based on a random sub-sample of WG sample, *n* = 500.

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

DIBELS Alphabetic Principle:
High Risk Not High Risk

Second End ORF Outcome:
Not High Risk

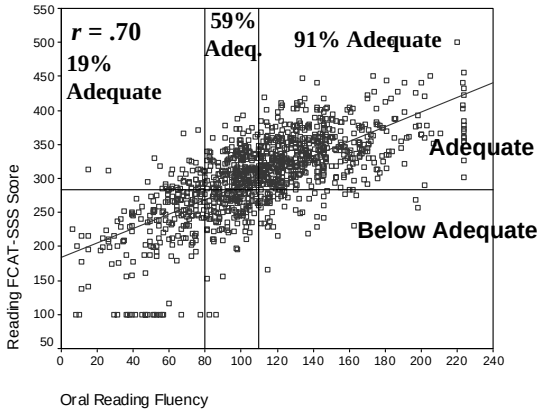
High Risk



- ***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

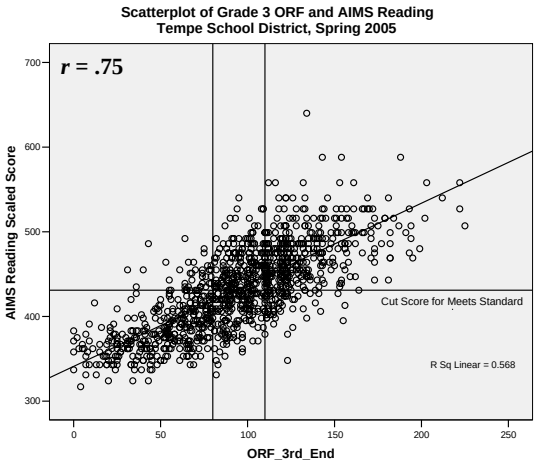
Between G3 ORF of 80 and 110, the odds are 59% the student will rank "adequate" on the FL State Assessment.



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.**

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

- The best reading assessments correlate in the range .60 to .80, consistent with the correlation of ORF and most other reading assessments.
- This means there will always be a zone of disagreement between any two criterion measures. How do we determine which assessment is the ***true*** gold standard assessment of reading outcomes?
- WRMT? NAEP? OSAT? SAT-10? FCAT? AIMS? DRA?



Wilson, unpublished data, 2005

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.

Educational Assessment is a Three-by-three World

First Grade End of Year NWF Initial Assessment:

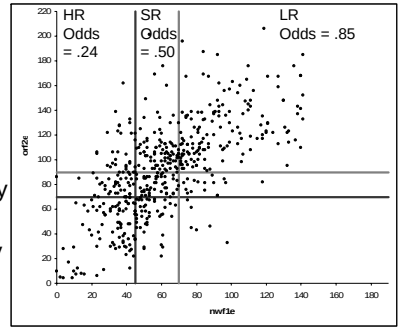
High Risk Some Risk Low Risk

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.

Applying Two-by-two Logic in a Three-by-three world

- Using 2 by 2 logic, 4 different sets of decision metrics 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.
- Sensitivity and Specificity depend on the cutpoint used on the screening assessment and on the outcome selected.

	LRD-LRO	LRD-HRO	HRD-LRO	HRD-HRO
True Negative	17089	19326	27830	35659
False Negative	2996	759	13609	5780
True Positive	23734	14307	13121	9286
False Positive	14992	24419	4251	8086
Sensitivity	0.89	0.95	0.49	0.62
Specificity	0.53	0.44	0.87	0.82
Negative Predictive Power	0.85	0.96	0.67	0.86
Positive Predictive Power	0.61	0.37	0.76	0.53
Accurate Classification	0.69	0.57	0.70	0.76
Decision Baserate	0.66	0.66	0.30	0.30

Note: Decision metrics based on Full WG sample, *n* = 58811.

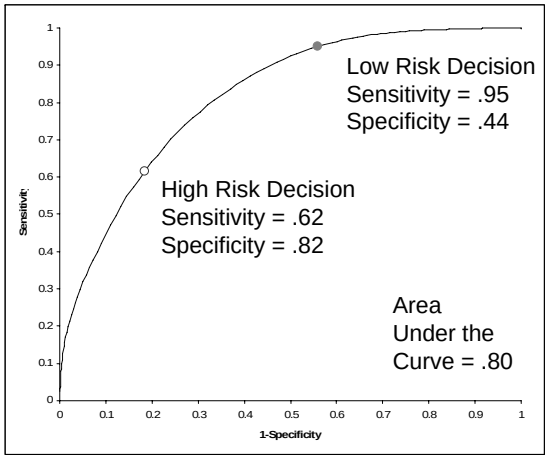
Part 3 Overview

- 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.
- Additional problems of Sensitivity and Specificity:
  - They depend on the choice of cutpoint. (ROC curves address the problem of different cutpoints, but ROC curves don't address the other problems of sensitivity and specificity).
  - They are affected by differences in baserate.
  - They are affected by differences in Tier 1 instruction and Tier 2 and Tier 3 interventions.
- Our recommendation:
  - Use likelihood or odds of achieving important educational outcomes to evaluate screening assessments.

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

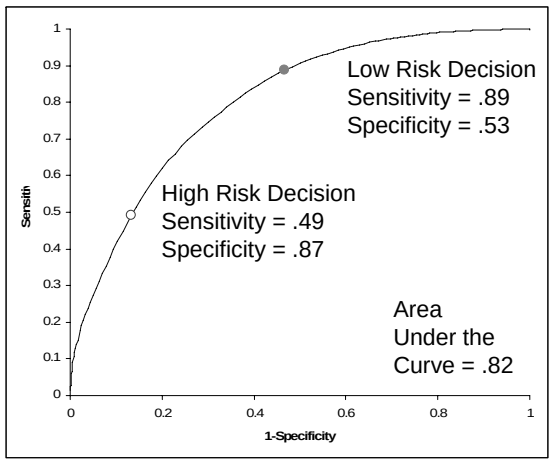
- It is meaningless to compare sensitivity indices on different tests (Swets, 1988) because:
  - Sensitivity *depends on the cutpoint for risk* that is selected. As we increase the cutpoint, sensitivity increases,
  - But, there is a trade-off. As we increase the cutpoint, the specificity decreases.
  - 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



Full WG Sample, *n* = 58811

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



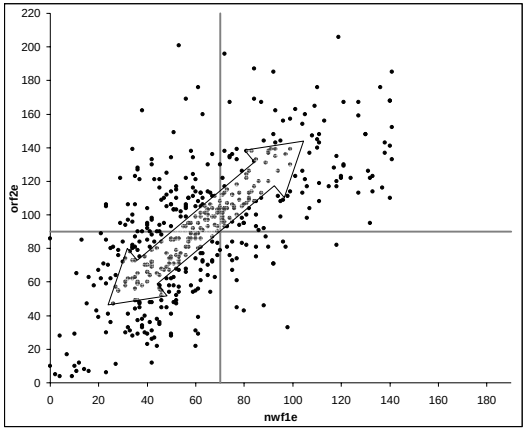
Full WG Sample, *n* = 58811

Part 3 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.
- Additional problems of Sensitivity and Specificity:
  - They depend on the choice of cutpoint. (ROC curves address the problem of different cutpoints, but ROC curves don't address the other problems of sensitivity and specificity).
  - They are affected by differences in baserate.
  - They are affected by differences in Tier 1 instruction and Tier 2 and Tier 3 interventions.
- Our recommendation:
  - Use likelihood or odds of achieving important educational outcomes to evaluate screening assessments.

The Problem of Differences in Baserates

Differences in baserate only do not change the nature of the underlying relation between the screener and the outcome.



- In a context with a greater (lesser) baserate of reading difficulty, more (less) students will be positive on the screener and more (less) students also will be positive on the outcome.
- The underlying relation between screener and outcome would remain the same, and students would move diagonally on the scatterplot.

Estimating Baserate

- In a setting like screening for tuberculosis, because the condition is truly present or absent at the time of screening and the outcome measure occurs before action or treatment,
  - Baserate is best estimated as the percent with a positive outcome on the criterion measure.
- In an educational setting, because the condition does not become present or absent until the outcome assessment and because the outcome is a joint result of initial skills and the instructional context,
  - Baserate is best estimated as the percent with a positive decision on the screening measure.

Relation of Baserate to Decision Metrics

- Screening decision baserate appears to be related to sensitivity and specificity in the mClass 137 district sample.
- A Monte Carlo study was conducted to examine the relation further. The 137 districts in the mClass sample were modeled with bivariate normal random data with identical differences in decision baserate but with no differences in instructional effectiveness.

	Correlation of index with screening decision baserate	
	mClass	Monte Carlo
LRD-LRO Sensitivity	.79	.83
LRD-LRO Specificity	-.93	-.95
LRO ROC Area Under the Curve (AUC)	-.41	-.12
LRD-LRO Positive Predictive Power	.40	.90
LRD-LRO Negative Predictive Power	-.21	-.80
LRD-LRO Classification Accuracy	-.20	.04
High Risk Decision Odds of Low Risk Outcome	-.10	-.74
Some Risk Decision Odds of Low Risk Outcome	-.13	-.79
Low Risk Decision Odds of Low Risk Outcome	-.21	-.80

Note. *n* = 137 districts.

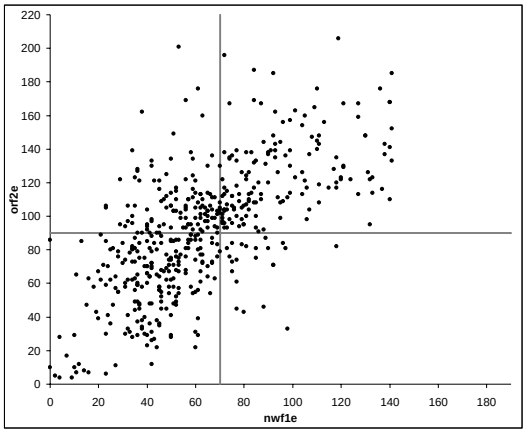
Although most indices are expected to be affected by intervening instruction, these metrics are likely to be the most sensitive to differences in Tier 1 instruction and Tier 2 and Tier 3 intervention.

Part 3 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.
- Additional problems of Sensitivity and Specificity:
  - They depend on the choice of cutpoint. (ROC curves address the problem of different cutpoints, but ROC curves don't address the other problems of sensitivity and specificity).
  - They are affected by differences in baserate.
  - They are affected by differences in Tier 1 instruction and Tier 2 and Tier 3 interventions.
- 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

Less Effective Tier 1 Instruction: As more students who screened low risk do not achieve the outcome, specificity and sensitivity decrease.

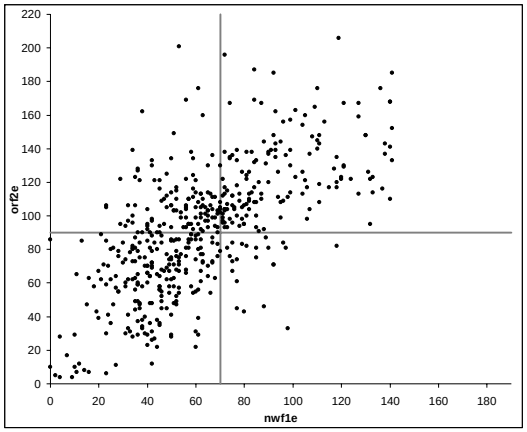


More Effective Tier 1 Instruction: As more students who screened low risk achieve the outcome, specificity and sensitivity increase.

- 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 <u>selected</u> students would move vertically on the scatterplot.

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

Less Effective Tier 2 & 3 Intervention: As more students who screened high or some risk do not achieve the outcome, specificity and sensitivity increase.



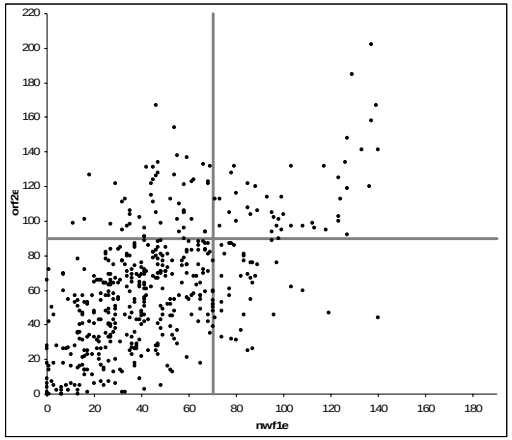
More Effective Tier 2 & 3 Intervention : As more students who screened high or some risk achieve the outcome, specificity and sensitivity decrease.

- 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 <u>selected</u> students would move vertically on the scatterplot.

The Big Ideas

- Differences in baserate **only** do not change the nature of the underlying relation between the screener and the outcome.
- 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 **<u>increases</u>** measures of sensitivity and specificity.
- Increasing the effectiveness of Tier 2 & 3 intervention **<u>decreases</u>** 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



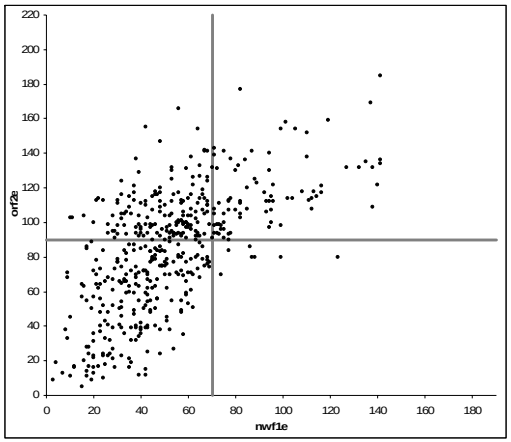
Sample District 1

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?

Note. Outcome baserate would be .80.

Sensitivity & Specificity Logic Doesn't Work



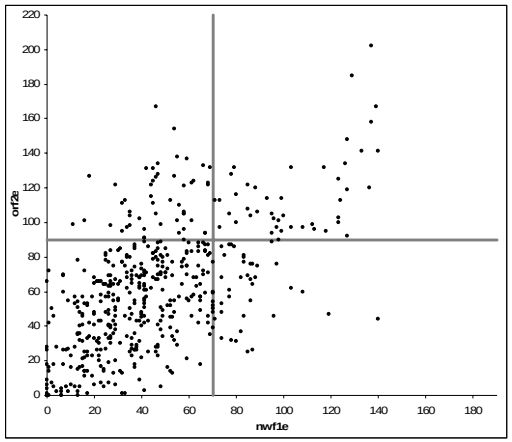
Sample District 2

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.

Sensitivity & Specificity Logic Doesn't Work



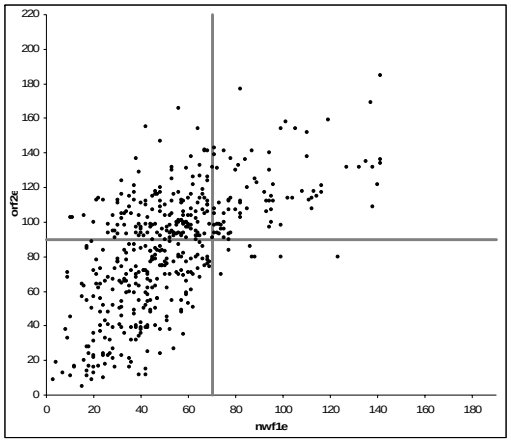
Sample District 1

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?

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 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.**

Note. Outcome baserate would be .49.

Part 3 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.
- Additional problems of Sensitivity and Specificity:
  - They depend on the choice of cutpoint. (ROC curves address the problem of different cutpoints, but ROC curves don't address the other problems of sensitivity and specificity).
  - They are affected by differences in baserate.
  - They are affected by differences in Tier 1 instruction and Tier 2 and Tier 3 interventions.
- 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.

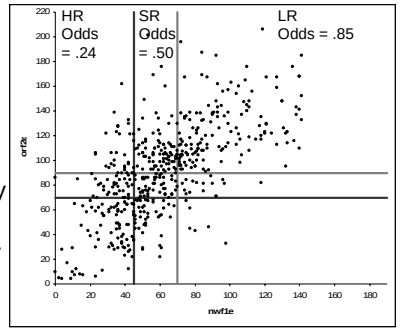
High Risk, Some Risk, and Low Risk Decisions

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

Second End of Year ORF Outcome:
Low Risk Reading Fluency

Some Risk Reading Fluency

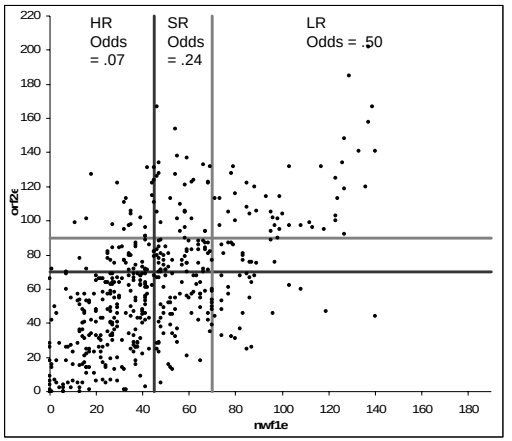
High Risk Reading Fluency



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

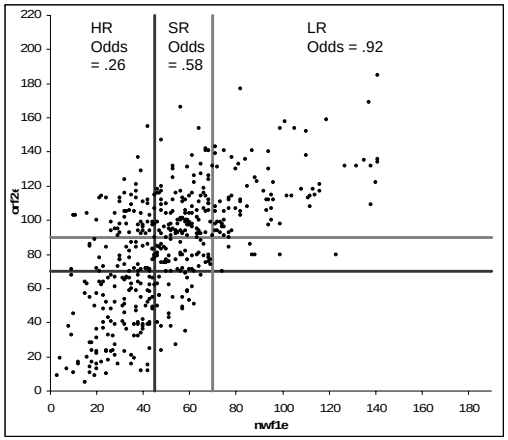
Note: Odds based on Full WG sample, *n* = 58811. Scatterplot based on a random sub-sample of WG sample, *n* = 500.

Sample District 1



- If fewer students with a low risk screening decision achieve the outcome than expected, we would want to examine instruction.
- If fewer students with a high risk or some risk screening decision achieve the outcome than expected, we would want to examine interventions.

Sample District 2



- If more students with a low risk screening decision achieve the outcome than expected, we would want to celebrate and maintain our instruction.
- If more students with a high risk or some risk screening decision achieve the outcome than expected, we would want to celebrate and maintain our interventions.

Decision Utility of DIBELS with the Full MClass Sample

Odds of Achieving ORF Benchmark Outcomes (Criterion)					
		G1 ORF EOY	G2 ORF BOY	G2 ORF MOY	G2 ORF EOY
Initial Support Decision Based on First Grade EOY NWF (Screen)	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

Educational Context is Fundamentally Different from Screening for TB

- Sensitivity and specificity may not make sense as primary evaluation metrics in an **educational context:**
  - The outcome is not a true dichotomous state that is present or absent. Reading skills are a continuum. There are a group of students we have reasonable agreement are on track for reading; a group of students we have reasonable agreement are not on track for reading; and a group of students whose reading status is uncertain:
    (a) we can't agree on a point that separates ok and not ok.
    (b) a student who is above any point on one test may not be above the corresponding point on a different measure.
    (c) students close to any point may be more similar than different.
  - There is not a gold standard for determining the true state. The true state is a value judgment that depends on measurement, social, and political context.
  - The true state does not exist at the time of screening, but becomes as a result of the effectiveness of instruction and intervention.
  - Whether the true state eventuates depends on the instructional context. The linkage/relation between screening (initial assessment) and outcome depends on the instructional context.
  - Treatment/action is or should be differentiated based upon student need.
  - False negatives should be minimized. False negatives are less desirable than false positives. We don't want to miss an opportunity to provide instruction to a student that helps put the odds in their favor of becoming a reader.

ROC Done Right

- Sensitivity and Specificity are fatally flawed as a means of evaluating an educational screening measure.
- ROC and area under the curve are interesting pieces of information to consider in evaluating the functioning of a screening measure in an educational context, but they may not be desirable as the primary consideration.
- A more desirable primary consideration may be the likelihood or odds, given initial skill level, of achieving an outcome where there is reasonable agreement the student has adequate reading skills.
- Screening Decision Baserate may affect all of the decision metrics.
- Evaluations or comparisons of instructional effectiveness may be most defensible when decision baserates are comparable or considered.