Classification Accuracy Update For Grades 6-8 SpringMath May 2026

Current SpringMath screening data have been evaluated in recent calls from the National Center for Intensive Intervention (NCII) for inclusion on their Tools' Chart. Those data can be viewed here https://charts.intensiveintervention.org/screening/tool/?id=5b5f5b465c8db3fd. This document will update existing data, previously reported to NCII, in an effort to continuously evaluate the accuracy of SpringMath screening and decision making.

Comparing these data to the thresholds used by the National Center for Intensive Intervention (NCII; www.intensiveintervention.org) for classification accuracy in MTSS, these data meet the highest standards used by NCII.

✓ Was an appropriate external measure of academic performance used as an outcome?
Yes. We used the Pennsylvania State System of Assessment (PSSA), which is the year-end accountability measure in Pennsylvania.

✓ Was risk adequately defined within an RTI approach to screening (i.e., 10th-20th percentile)?
Yes, we used below the 20th percentile on PSSA, as preferred by NCII. We also report accuracies using proficient/nonproficient on the PSSA, as is generally preferred by systems.

✓ Were the classification analyses and cut-points adequately performed?
Yes. We used the methodology that we have often provided in peer-reviewed published studies.

✓ The lower bound of the confidence interval around the Area Under the Curve (AUC) estimate ≥ 0.80 and
Lower bound of AUCs for all grades and all analyses are less than or equal to 0.80.

✓ Sensitivity ≥ 0.80 and Specificity ≥ 0.80
Sensitivity and specificity are equal to or greater than 0.80 for all grades and all analyses.

These data were collected from a district in the northeastern U.S. for students in grades 6 through 8. Demographic data were available for all grades. Gender was roughly evenly divided, with slightly more females than males in Grade 8 which was also a smaller sample. Most participants were White. Eight to twenty-one percent of the sample received special education services and 1% of students in Grade 8 were considered English Language Learners. The samples for the fall and winter analyses were very similar and full details are provided for both samples in the following tables.

Sample Details for Fall

Sample	Grade 6	Grade 7	Grade 8
Sample Size	171	198	50
Geographic Representation	Northeast (PA)	Northeast (PA)	Northeast (PA)
Male	86 (50%)	97 (49%)	18 (36%)
Female	85 (50%)	101 (51%)	32 (64%)
Other	0	0	0
Gender Unknown	0	0	0
White, Non-Hispanic	159 (93%)	185 (93%)	49 (98%)
Black, Non-Hispanic	4 (2%)	4 (2%)	0
Hispanic	1 (1%)	4 (2%)	0
Asian/Pacific Islander	0	0	1 (2%)
American Indian/Alaska Native	0	0	0
Other/Multiple	7 (4%)	5 (3%)	0
Race / Ethnicity Unknown	0	0	0
Low SES	No Data	No Data	No Data
IEP or diagnosed disability	30 (18%)	42 (21%)	4 (8%)
English Language Learner	0	0	1 (2%)

* Fall 2024 (screening), state assessment 2025 (criterion); Percents may not sum exactly to 100% due to rounding.

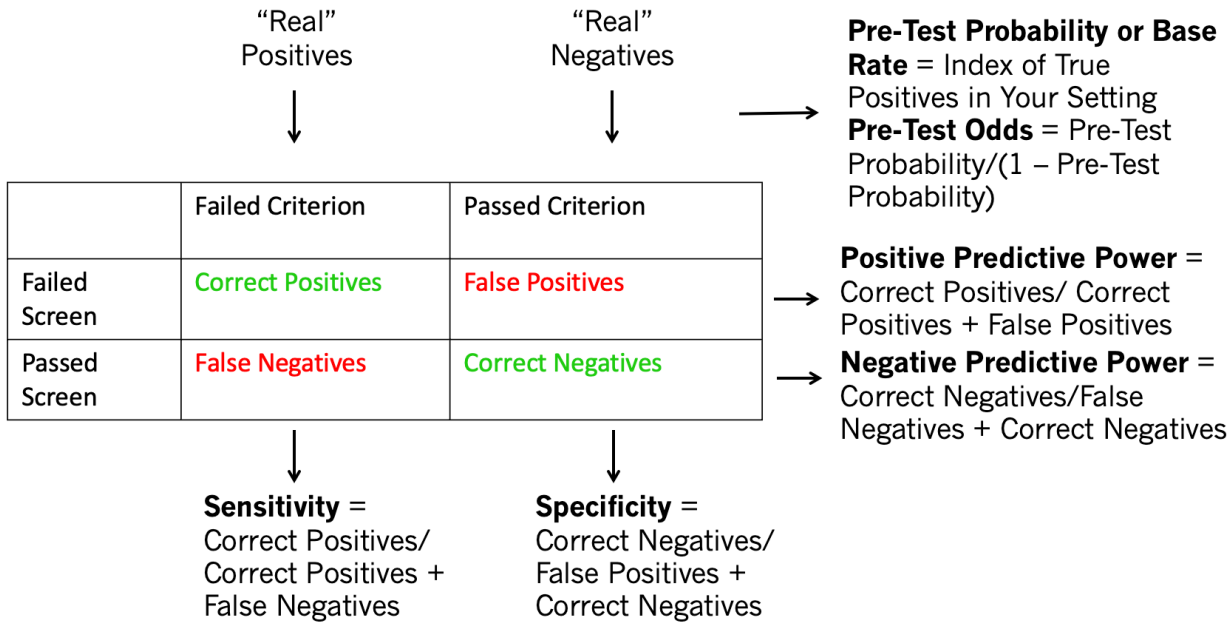
Sample Details for Winter

	Grade 6	Grade 7	Grade 8
Sample Size	119	193	71
Geographic Representation	Northeast (PA)	Northeast (PA)	Northeast (PA)
Male	55 (46%)	100 (52%)	29 (41%)
Female	64 (54%)	93 (48%)	42 (59%)
Other	0	0	0
Gender Unknown	0	0	0
White, Non-Hispanic	114 (96%)	182 (94%)	66 (93%)
Black, Non-Hispanic	4 (3%)	3 (2%)	1 (1%)
Hispanic	0	4 (2%)	1 (1%)
Asian/Pacific Islander	0	0	1 (1%)
American Indian/Alaska Native	0	0	0
Other/Multiple	1 (0.8%)	4 (2%)	2 (3%)
Race / Ethnicity Unknown	0	0	0
Low SES	No Data	No Data	No Data
IEP or diagnosed disability	18 (15%)	48 (25%)	12 (17%)
English Language Learner	0	0	1 (1%)

* Winter 2024 (screening), state assessment 2025 (criterion); Percents may not sum exactly to 100% due to rounding.

Previous research has demonstrated reliability of scores obtained on SpringMath measures and demonstrated a lack of bias in the scores. Details of past research studies examining the technical adequacy of SpringMath measures can be viewed here: https://springmath.org/ebook. This update will focus only on classification accuracy data.

In screening, the most critical validity evidence is classification accuracy. It is not possible to attain adequate thresholds of classification accuracy in the absence of well-correlated scores between predictor and criterion, but correlation alone does not guarantee adequate classification accuracy. In classification analysis research, a rule is applied to determine risk from the screening scores. Scores below the rule are considered at-risk (screening-positive) and scores above the rule are considered not at-risk (screening-negative). Those children also have scores on a reference criterion for which a different rule is applied coding those students as scoring nonproficient on the criterion (criterion-positive) or proficient on the criterion (criterion-negative). These coding procedures result in a four-cell contingency table that allows us to characterize the accuracy of decisions based on the screenings, using the standard classification agreement metrics, the most essential of which are sensitivity (the capacity of the screening to detect true positives or students who will score nonproficient on the year-end test) and specificity (the capacity of the screening to detect true negatives or students who will score proficient on the year-end test). There is always a trade-off between sensitivity (avoiding false negative errors by using a more liberal threshold for screening risk) and specificity (avoiding false positives by using a more stringent threshold for screening risk). This trade-off is unavoidable and the Area Under the Curve from the Receiver Operating Curve makes this trade-off apparent while illustrating the value of the screening in separating true positives from true negatives across the full range of possible screening scores.



The criterion used to evaluate screening accuracy is generally year-end state test performance when those data are available.

In the sections that follow, we provide updated classification accuracy data for fall screenings and winter screenings in grades 6-8 in SpringMath. We also examine the classification accuracy of classwide intervention risk since classwide intervention is the second screening gate in actual SpringMath implementation.

We report accuracy against the Pennsylvania System of School Assessment (PSSA) for which we evaluate on two thresholds: scoring below the 20th percentile and scoring nonproficient.

In all cases, at every grade level, sensitivity and specificity values equal or exceed .80 as does the lower bound of the confidence interval around the Receiver Operating Curve (ROC).

Scatterplots showing screening accuracy and Receiver Operating Curves are appended at the end of this document for the fall screenings (winter screenings were nearly identical).

Fall Screenings Grades 6-8 and Year-End State Test Scores

	Grade 6	Grade 7	Grade 8
Cut point: Criterion measure*	907	865	891
Cut point: Screening measure	12	5	17
Base rate	0.20	0.20	0.20
False positive rate	0.20	0.15	0.05
False negative rate	0.14	0.13	0.20
Sensitivity	0.86	0.88	0.80
Specificity	0.81	0.85	0.95
Positive predictive power	0.60	0.60	0.80
Negative predictive power	0.95	0.96	0.95
AUC (95% CI)	0.87 (0.80 - 0.92)	0.92 (0.86 - 0.95)	0.96 (0.86 - 0.99)
Correct positives (CP)	37	35	8
False positives (FP)	25	23	2
False negatives (FN)	6	5	2
Correct negatives (CN)	103	135	38
Total (CP + FP + FN + CN)	171	198	50
Criterion-positive (CP + FN)	43	40	10
Flagged by screener (CP + FP)	62	58	10

*Criterion in grades 6-8 is the total mathematics scaled score for the Pennsylvania System of School Assessment (PSSA). The criterion reference group is calculated as below the 20th percentile on the PSSA in the analysis sample, consistent with the preferred methodology of the National Center for Intensive Intervention.

Fall Screenings Grades 6-8 and Year-End State Test Proficiency

Educational leaders generally want to predict who is going to score proficient and who is going to score non-proficient on the year-end state test. Thus, on the same sample of grade 6-8 students, we also examined screening accuracy in predicting proficiency on the year-end state test (PSSA). Accuracies tracked very closely to predicting which students would score below the 20th percentile on the state test. In all grades, sensitivity and specificity met or exceeded the most rigorous thresholds of 0.80, as did the lower bound of the confidence interval for the Area Under the Curve from the Receiver Operating Curve (ROC) Analysis.

	Grade 6	Grade 7	Grade 8
Cut point: Criterion measure*	1000	1000	1000
Cut point: Screening measure	14	10	32
Base rate	0.53	0.58	0.66
False positive rate	0.17	0.19	0.13
False negative rate	0.17	0.18	0.18
Sensitivity	0.83	0.82	0.82
Specificity	0.83	0.82	0.88
Positive predictive power	0.82	0.84	0.93
Negative predictive power	0.84	0.80	0.70
AUC (95% CI)	0.92 (0.88 - 0.95)	0.92 (0.87 - 0.95)	0.91 (0.80 - 0.97)
Correct positives (CP)	68	87	28
False positives (FP)	15	17	2
False negatives (FN)	14	19	6
Correct negatives (CN)	74	75	14
Total (CP + FP + FN + CN)	171	198	50
Criterion-positive (CP + FN)	82	106	34
Flagged by screener (CP + FP)	83	104	30

*The reference criterion in these data was proficient or non-proficient on the year-end state test in Pennsylvania (PSSA). We report the accuracies relative to proficiency because typically district leaders want to predict whether students will score proficient or not on the year-end test.

Winter Screenings Grades 6-8 and Year-End State Test Scores

	Grade 6	Grade 7	Grade 8
Cut point: Criterion measure*	907	865	855
Cut point: Screening measure	21	21	13
Base rate	0.20	0.20	0.20
False positive rate	0.20	0.20	0.18
False negative rate	0.12	0.20	0.19
Sensitivity	0.89	0.80	0.81
Specificity	0.80	0.80	0.82
Positive predictive power	0.55	0.52	0.57
Negative predictive power	0.96	0.94	0.94
AUC (95% CI)	0.89 (0.82 - 0.94)	0.89 (0.83 - 0.93)	0.91 (0.80 - 0.96)
Correct Positives (CP)	23	32	13
False Positives (FP)	19	30	10
False Negatives (FN)	3	8	3
Correct Negatives (CN)	74	123	45
Total (CP + FP + FN + CN)	119	193	71
Criterion-positive (CP + FN)	26	40	16
Flagged by screener (CP + FP)	42	62	23

*Criterion in grades 6-8 is the total mathematics scaled score for the Pennsylvania System of School Assessment (PSSA). The criterion reference group is calculated as below the 20th percentile on the PSSA in the analysis sample, consistent with the preferred methodology of the National Center for Intensive Intervention.

Winter Screenings Grades 6-8 and Year-End State Test Proficiency

Educational leaders generally want to predict who is going to score proficient and who is going to score non-proficient on the year-end state test. Thus, on the same sample of grade 6-8 students, we also examined screening accuracy in predicting proficiency on the year-end state test (PSSA). Accuracies tracked very closely to predicting which students would score below the 20th percentile on the state test. In all grades, sensitivity and specificity met or exceeded the most rigorous thresholds of 0.80.

	Grade 6	Grade 7	Grade 8
Cut point: Criterion measure*	1000	1000	1000
Cut point: Screening measure	22	28	26
Base rate	0.53	0.58	0.66
False positive rate	0.10	0.15	0.14
False negative rate	0.19	0.20	0.14
Sensitivity	0.81	0.80	0.86
Specificity	0.90	0.85	0.86
Positive predictive power	0.84	0.87	0.93
Negative predictive power	0.88	0.77	0.73
AUC (95% CI)	0.92 (0.84 - 0.96)	0.90 (0.85 - 0.93)	0.93 (0.84 - 0.97)
Correct positives (CP)	38	86	42
False positives (FP)	7	13	3
False negatives (FN)	9	22	7
Correct negatives (CN)	65	72	19
Total (CP + FP + FN + CN)	119	193	71
Criterion-positive (CP + FN)	47	108	49
Flagged by screener (CP + FP)	45	99	45

*The reference criterion in these data was proficient or non-proficient on the year-end state test in Pennsylvania (PSSA). We report the accuracies relative to proficiency because typically district leaders want to predict whether students will score proficient or not on the year-end test.

Classwide Intervention Risk Grades 6-8 and Year-End State Test

In SpringMath, classwide math intervention is used as the second screening gate. Thus, it can and should be evaluated for accuracy as a screening mechanism. We have previously reported accuracy data for classwide math intervention as a screening mechanism on the NCII tools' chart meeting the criteria to earn full bubbles at some grade levels. Data are presented below for classwide intervention as a screening for all grades K-5. The proportion of opportunities to be at risk that a given student met the risk criterion was the predictor score. The reference criterion used

below is scoring below the 20th percentile on the year-end state test in Pennsylvania (PSSA) in Grades 6-8.

	Grade 6	Grade 7	Grade 8
Cut point: Criterion*	916	932	855
Cut point: Classwide Intervention Risk	0.082	0.082	0.203
Sensitivity	0.85	0.89	0.80
Specificity	0.84	0.88	0.94
Positive predictive power	0.71	0.70	0.57
Negative predictive power	0.92	0.96	0.98
AUC (95% CI)	0.89 (0.80- 0.94)	0.91 (0.80- 0.96)	0.94 (0.80 - 0.98)
Correct positives	34	23	4
False positives	14	10	3
False negatives	6	3	1
Correct negatives	71	73	49

*Criterion in grades 6-8 is the total mathematics scaled score for the Pennsylvania System of School Assessment (PSSA). The criterion reference group is calculated as below the 20th percentile on the PSSA in the analysis sample, consistent with the preferred methodology of the National Center for Intensive Intervention. Classwide risk is the proportion of opportunities to be at risk during classwide intervention for which a student met the risk criterion to advance to intensified instruction.

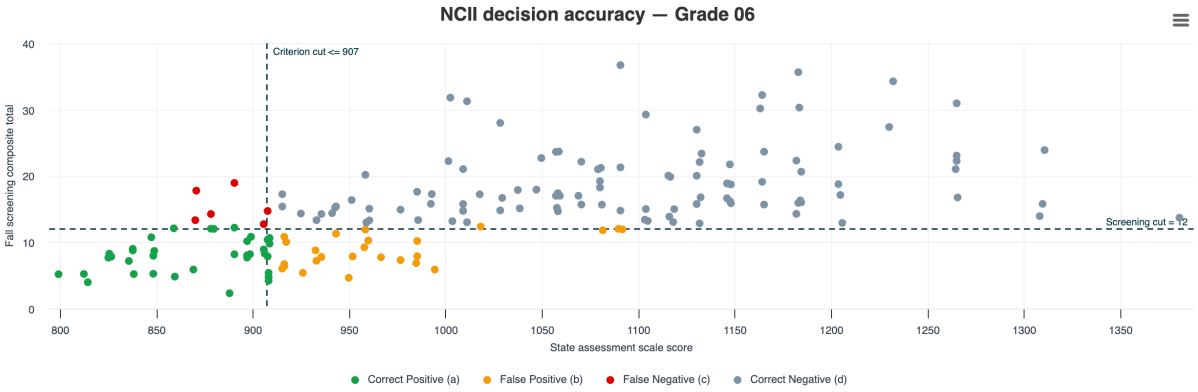
Classwide Intervention Risk in Grades 6-8 and Year-End State Test Proficiency

Because leaders are typically interested in predicting proficiency, the table below reports the accuracy of classwide intervention risk as a second screening gate in predicting proficiency on the year-end state test in math in Pennsylvania) for grades 6-8.

	Grade 6	Grade 7	Grade 8
Cut point: Criterion*	1000	1000	1000
Cut point: Classwide Intervention Risk	0.042	0.04	0.036
Sensitivity	0.84	0.82	0.92
Specificity	0.80	0.86	0.88
Positive predictive power	0.86	0.84	0.85
Negative predictive power	0.77	0.85	0.93
AUC (95% CI)	0.88 (0.82 - 0.93)	0.90 (0.82 - 0.95)	0.92 (0.80 - 0.97)
Correct positives	63	41	23
False positives	10	8	4
False negatives	12	9	2
Correct negatives	40	51	28

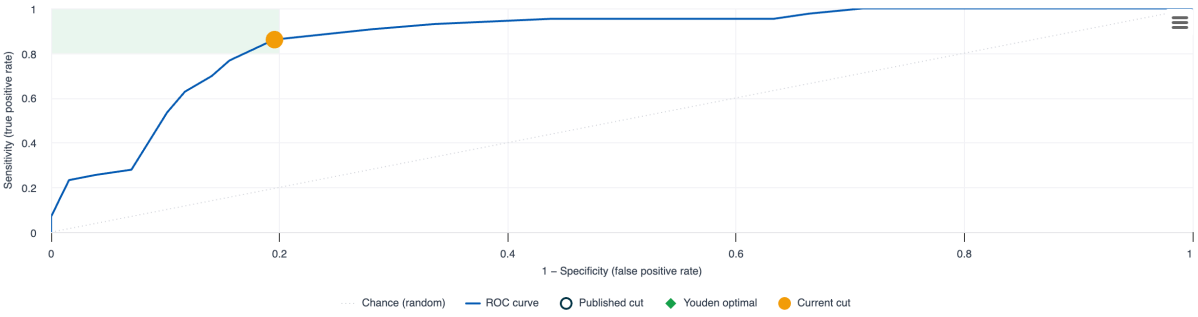
*Criterion in grades 6-8 is the total mathematics scaled score for the Pennsylvania System of School Assessment (PSSA). The criterion reference group is calculated as proficient/nonproficient on the PSSA. Classwide risk is the proportion of opportunities to be at risk during classwide intervention for which a student met the risk criterion to advance to intensified instruction.

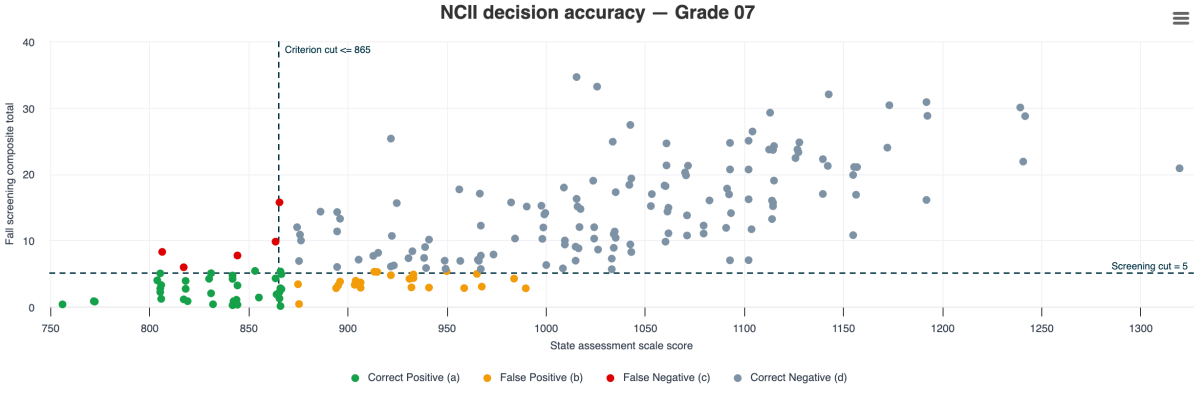
Decision Accuracies and Receiver Operating Curves for Fall Screenings



ROC curve — grade 06

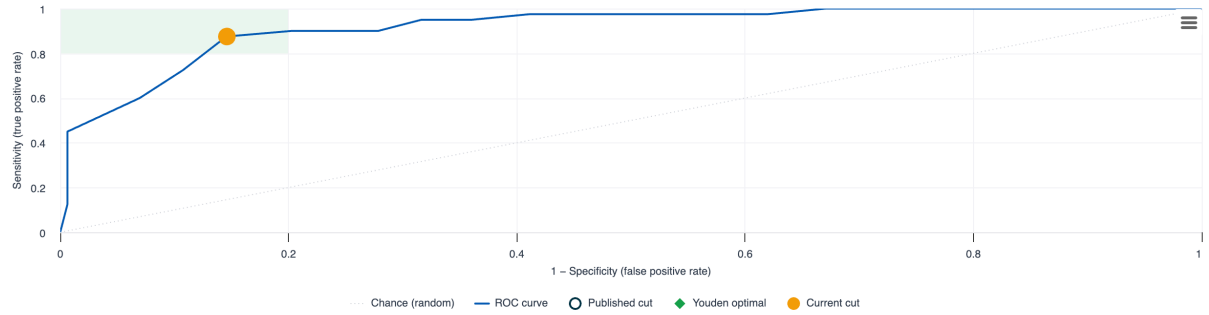
Sensitivity vs. 1 − specificity across every possible screener cut. The faint green band marks the NCII Full Bubble qualifying region (sens ≥ 0.80, FPR ≤ 0.20). Drag the open circle to test a hypothetical cut — the analysis-table row above will recompute its CP/FP/FN/CN, sens, spec, PPV, and NPV. AUC is threshold-independent and won't change.

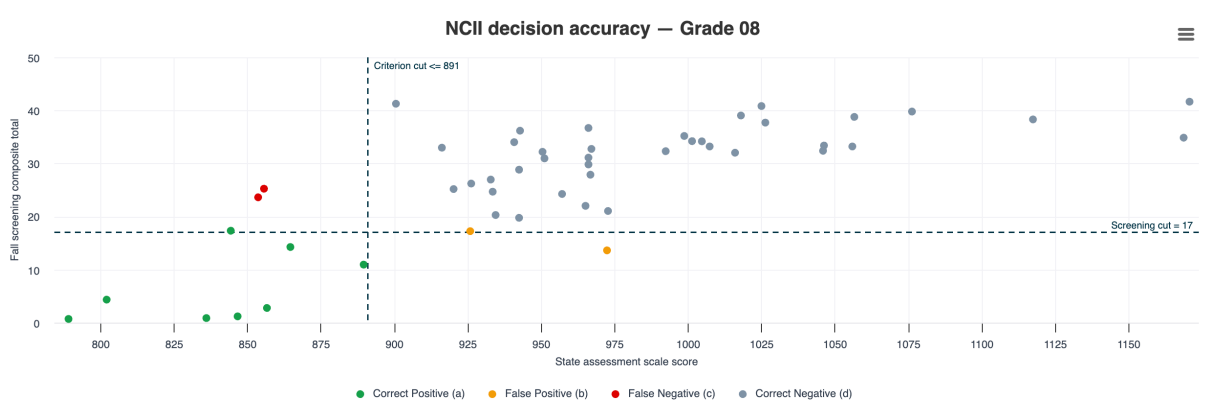




ROC curve — grade 07

Sensitivity vs. 1 − specificity across every possible screener cut. The faint green band marks the NCII Full Bubble qualifying region (sens ≥ 0.80, FPR ≤ 0.20). Drag the open circle to test a hypothetical cut — the analysis-table row above will recompute its CP/FP/FN/CN, sens, spec, PPV, and NPV. AUC is threshold-independent and won't change.





ROC curve — grade 08

Sensitivity vs. 1 − specificity across every possible screener cut. The faint green band marks the NCII Full Bubble qualifying region (sens ≥ 0.80, FPR ≤ 0.20). Drag the open circle to test a hypothetical cut — the analysis-table row above will recompute its CP/FP/FN/CN, sens, spec, PPV, and NPV. AUC is threshold-independent and won't change.