Classification Accuracy Update 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.

These data were collected from a district in the southwestern U.S. for students in Kindergarten through Grade 5. Demographic data were available for students in grades 3-5 which was consistent with school-level demographic data and thus representative of the sample. Gender was roughly evenly divided and most participants were either Hispanic or White, each of which accounted for roughly half of 90% of the sample. The remaining 10% of the sample self-identified as Black, Multiple, Asian, or American Indian/Alaska Native. Fourteen to eighteen percent of the sample received special education services and 1 to 3% of students 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

	Grade 3	Grade 4	Grade 5
Sample Size	360	241	622
Geographic Representation	Mountain (AZ)	Mountain (AZ)	Mountain (AZ)
Male	175 (49%)	133 (55%)	310 (50%)
Female	185 (51%)	108 (45%)	312 (50%)
Other	0	0	0
Gender Unknown	0	0	0
White, Non-Hispanic	159 (44%)	113 (47%)	280 (45%)
Black, Non-Hispanic	16 (4%)	5 (2%)	29 (5%)
Hispanic	149 (41%)	104 (43%)	240 (39%)
Asian/Pacific Islander	3 (0.8%)	6 (3%)	11 (2%)
American Indian/Alaska Native	1 (0.3%)	1 (0.4%)	1 (0.2%)
Other/Multiple	32 (9%)	12 (5%)	61 (10%)
Race / Ethnicity Unknown	0	0	0
Low SES	No Data	No Data	No Data
IEP or diagnosed disability	59 (16%)	44 (18%)	91 (14%)
English Language Learner	11 (3%)	3 (1%)	12 (2%)

* Fall 2024 (screening), state assessment 2025 (criterion); Percents may not sum exactly to 100% due to rounding. Data not available for Grades K-2.

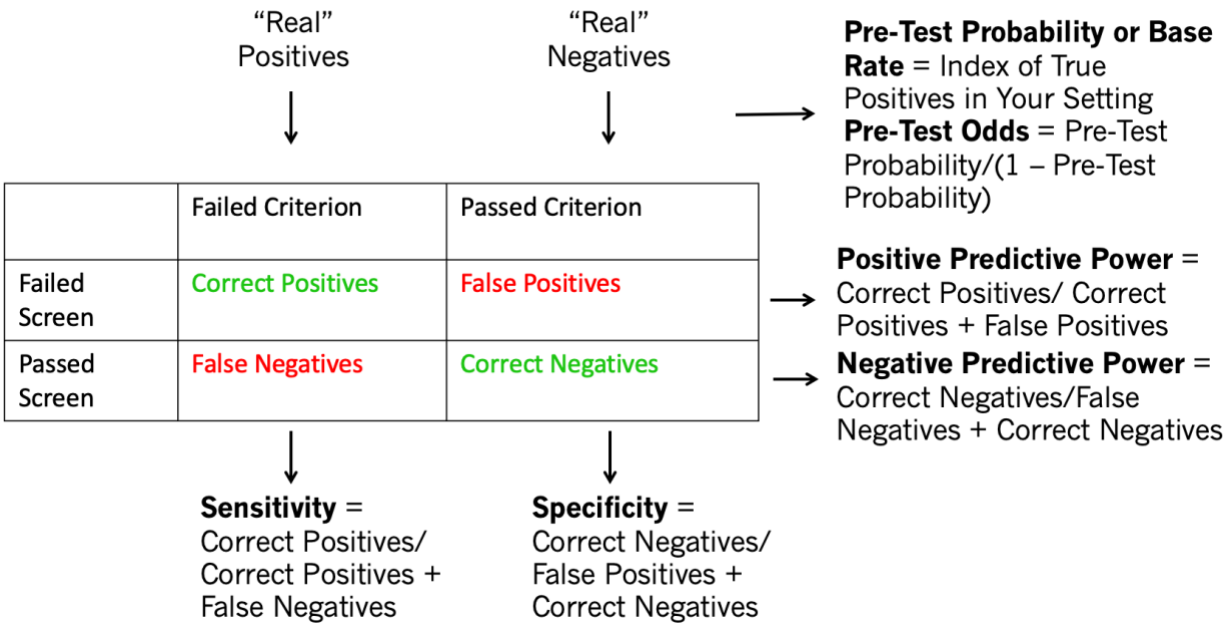
Sample Details for Winter

	Grade 3	Grade 4	Grade 5
Sample Size	374	242	642
Geographic Representation	Mountain (AZ)	Mountain (AZ)	Mountain (AZ)
Male	182 (49%)	132 (55%)	324 (51%)
Female	192 (51%)	110 (46%)	318 (50%)
Other	0	0	0
Gender Unknown	0	0	0
White, Non-Hispanic	161 (43%)	112 (46%)	286 (45%)
Black, Non-Hispanic	16 (4%)	6 (3%)	30 (5%)
Hispanic	158 (42%)	106 (44%)	253 (39%)
Asian/Pacific Islander	3 (0.8%)	5 (2%)	11 (2%)
American Indian/Alaska Native	1 (0.3%)	0	1 (0.2%)
Other/Multiple	35 (9%)	13 (5%)	61 (10%)
Race / Ethnicity Unknown	0	0	0
Low SES	No Data	No Data	No Data
IEP or diagnosed disability	60 (16%)	43 (18%)	94 (15%)
English Language Learner	15 (4%)	4 (2%)	15 (2%)

*Winter 2024 (screening), state assessment 2025 (criterion); Percents may not sum exactly to 100% due to rounding. Data not available for Grades K-2.

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. For younger students, the criterion assessment has to be justified. In our technical adequacy data we use a composite score from three to four administered curriculum-based measures in the spring for students in kindergarten through grade 2.

In the sections that follow, we provide updated classification accuracy data for fall screenings and winter screenings in grades K-5 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 composite score performance K-2, and on the Arizona Academic Standards Assessment for which we evaluate on two thresholds-- scoring below the 20th percentile on the AASA and scoring nonproficient on the AASA.

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

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 K-5 with Year-End Composite or State Test Scores

	Kindergarten	Grade 1	Grade 2	Grade 3	Grade 4	Grade 5
Criterion	Composite	Composite	Composite	ASSA	ASSA	ASSA
Cut point: Criterion measure	35	60	10	3503	3534	3577
Cut point: Screening measure	12	60	28	20	62	26
Base rate	0.20	0.20	0.20	0.20	0.20	0.20
False positive rate	0.20	0.17	0.18	0.15	0.02	0.11
False negative rate	0.19	0.12	0.19	0.13	0.18	0.20
Sensitivity	0.81	0.88	0.81	0.87	0.82	0.80
Specificity	0.80	0.83	0.82	0.85	0.98	0.90
Positive predictive power	0.51	0.57	0.53	0.60	0.91	0.67
Negative predictive power	0.94	0.97	0.95	0.96	0.95	0.94
AUC (95% CI)	0.88 (0.85 - 0.91)	0.89 (0.83 - 0.93)	0.90 (0.87 - 0.92)	0.94 (0.91 - 0.96)	0.96 (0.92 - 0.98)	0.91 (0.87 - 0.94)
Correct Positives (CP)	94	44	110	65	41	105
False Positives (FP)	92	33	97	44	4	52
False Negatives (FN)	22	6	26	10	9	26
Correct Negatives (CN)	367	165	445	241	187	439
Total (CP + FP + FN + CN)	575	248	678	360	241	622
Criterion-positive (CP + FN)	116	50	136	75	50	131
Flagged by screener (CP + FP)	186	77	207	109	45	157

*Criterion in grades 3-5 is the total mathematics scaled score for the Arizona Academic Standards Assessment (ASSA). The criterion reference group is calculated as below the 20th percentile on the ASSA in the analysis sample, consistent with the preferred methodology of the National Center for Intensive Intervention (NCII).

Fall Screenings Grades 3-5 with 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 3-5 students for whom year-end state test data are available, we also examined screening accuracy in predicting proficiency on the year-end state test (ASSA). Accuracies tracked very closely to predicting which students would score below the 20th percentile on the state test. In all grades, sensitivity and specificity 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 Analysis.

	Grade 3	Grade 4	Grade 5
Criterion	ASSA Proficiency	ASSA Proficiency	ASSA Proficiency
Cut point: Criterion measure	3531	3562	3595
Cut point: Screening measure	22	81	32
Base rate	.35	.34	.32
False positive rate	0.15	0.16	0.16
False negative rate	0.20	0.19	0.20
Sensitivity	0.80	0.81	0.80
Specificity	0.85	0.84	0.84
Positive predictive power	0.72	0.69	0.67
Negative predictive power	0.9	0.91	0.91
AUC (95% CI)	0.91 (0.87 - 0.94)	0.91 (0.86 - 0.94)	0.90 (0.87 - 0.93)
Correct Positives (CP)	94	61	144
False Positives (FP)	36	27	70
False Negatives (FN)	23	14	36
Correct Negatives (CN)	207	139	372
Total (CP + FP + FN + CN)	360	241	622
Criterion-positive (CP + FN)	117	75	180
Flagged by screener (CP + FP)	130	88	214

*The reference criterion in these data was proficient or non-proficient on the year-end state test in Arizona (ASSA). 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 K-5 and Year-End Composite or State Test Scores

	Kindergarten	Grade 1	Grade 2	Grade 3	Grade 4	Grade 5
Criterion	Composite	Composite	Composite	ASSA	ASSA	ASSA
Cut point: Criterion measure	36	59	10	3503	3534	3577
Cut point: Screening measure	27	39	48	27	23	57
Base rate	0.20	0.20	0.20	0.20	0.20	0.20
False positive rate	0.16	0.12	0.20	0.18	0.19	0.19
False negative rate	0.17	0.15	0.20	0.11	0.20	0.20
Sensitivity	0.83	0.85	0.80	0.89	0.80	0.80
Specificity	0.84	0.89	0.80	0.82	0.81	0.81
Positive predictive power	0.58	0.65	0.50	0.57	0.53	0.53
Negative predictive power	0.95	0.96	0.94	0.96	0.94	0.94
AUC (95% CI)	0.92 (0.89 - 0.94)	0.94 (0.91 - 0.96)	0.90 (0.86 - 0.92)	0.94 (0.90 - 0.96)	0.90 (0.85 - 0.94)	0.89 (0.85 - 0.92)
Correct Positives (CP)	102	45	117	70	41	109
False Positives (FP)	75	24	115	52	36	96
False Negatives (FN)	21	8	29	9	10	28
Correct Negatives (CN)	399	185	448	243	155	409
Total (CP + FP + FN + CN)	597	262	709	374	242	642
Criterion-positive (CP + FN)	123	53	146	79	51	137
Flagged by screener (CP + FP)	177	69	232	122	77	205

*Criterion in grades 3-5 is the total mathematics scaled score for the Arizona Academic Standards Assessment (ASSA). The criterion reference group is calculated as below the 20th percentile on the ASSA in the analysis sample, consistent with the preferred methodology of the National Center for Intensive Intervention.

Winter Screenings Grades 3-5 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 3-5 students for whom year-end state test data are available, we also examined screening accuracy in predicting proficiency on the year-end state test (ASSA). Accuracies tracked very closely to predicting which students would score below the 20th percentile on the state test. In all grades, sensitivity and specificity exceeded the most rigorous thresholds of 0.80.

	Grade 3	Grade 4	Grade 5
Criterion	ASSA Proficiency	ASSA Proficiency	ASSA Proficiency
Cut point: Criterion measure	3531	3562	3595
Cut point: Screening measure	31	25	62
Base rate	.35	.34	.32
False positive rate	0.19	0.20	0.20
False negative rate	0.20	0.20	0.20
Sensitivity	0.80	0.80	0.80
Specificity	0.81	0.80	0.80
Positive predictive power	0.68	0.64	0.63
Negative predictive power	0.89	0.9	0.91
AUC (95% CI)	0.89 (0.85 - 0.92)	0.87 (0.82 - 0.91)	0.88 (0.85 - 0.91)
Correct Positives (CP)	99	59	152
False Positives (FP)	47	33	91
False Negatives (FN)	25	15	38
Correct Negatives (CN)	203	135	361
Total (CP + FP + FN + CN)	374	242	642
Criterion-positive (CP + FN)	124	74	190
Flagged by screener (CP + FP)	146	92	243

*The reference criterion in these data was proficient or non-proficient on the year-end state test in Arizona (ASSA). 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 Risk Grades K-5 and Composite or 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 spring composite in Kindergarten through Grade 2 and scoring below the 20th percentile on the year-end state test in Arizona (AASA) in Grades 3-5.

	Kindergarten	Grade 1	Grade 2	Grade 3	Grade 4	Grade 5
Classwide Intervention Risk	0.121	0.14286	0.21429	0.08333	0.11111	0.087
Criterion cut*	36	52	11	3513	3530	3584
Sensitivity	0.87	0.86	0.82	0.90	0.87	0.86
Specificity	0.85	0.82	0.88	0.82	0.82	0.85
Positive predictive power	0.50	0.48	0.48	0.65	0.50	0.54
Negative predictive power	0.98	0.97	0.97	0.96	0.97	0.97
AUC (95% CI)	0.90 (0.84 - 0.94)	0.89 (0.80 - 0.94)	0.92 (0.87 - 0.94)	0.88 (0.82 - 0.92)	0.88 (0.80 - 0.93)	0.89 (0.82 - 0.93)
Correct positives	54	30	45	53	33	44
False positives	54	32	49	28	33	38
False negatives	8	5	10	6	5	7
Correct negatives	316	144	361	130	155	215

*The criterion for these analyses was scoring below the 20th percentile on the spring composite score in grades K-2 or below the 20th percentile on the AASA in grades 3-5.

In all grades, classwide intervention as a screening mechanism exceeded sensitivity and specificity of .80. The lower bound of the Area Under the Curve (AUC) estimates from the ROC analyses also exceeded .80 in every grade.

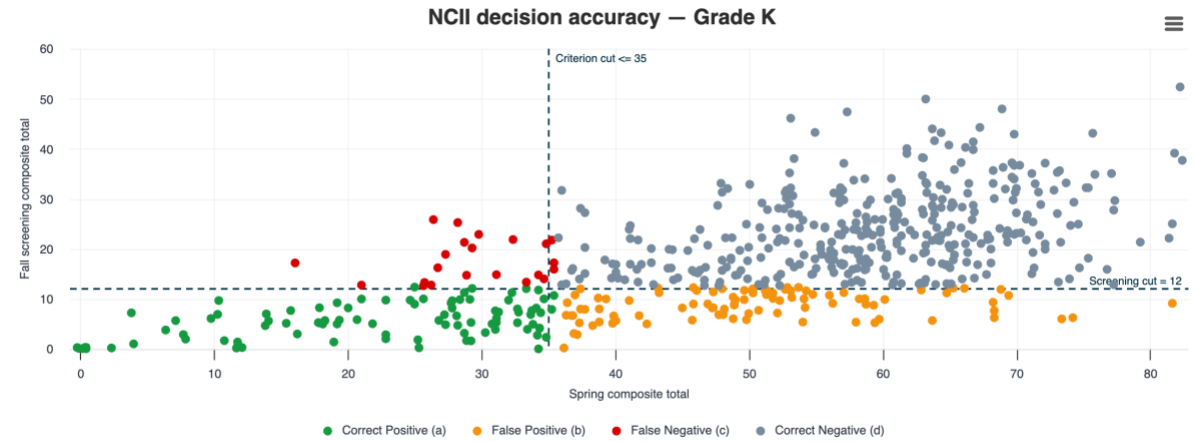
Classwide Intervention in Grades 3-5 and Year-End State Test Proficiency

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 Arizona (AASA) for grades 3-5.

	Grade 3	Grade 4	Grade 5
Classwide Intervention Risk	0.042	0.026	0.05882
Criterion cut*	3531	3562	3595
Sensitivity	0.80	0.80	0.83
Specificity	0.89	0.86	0.82
Positive predictive power	0.82	0.75	0.56
Negative predictive power	0.88	0.89	0.95
AUC (95% CI)	0.85 (0.80 - 0.90)	0.86 (0.80 - 0.90)	0.86 (0.80 - 0.91)
Correct positives	66	63	55
False positives	15	21	44
False negatives	17	16	11
Correct negatives	119	126	194

*The criterion for these analyses was proficiency on the AASA for mathematics.

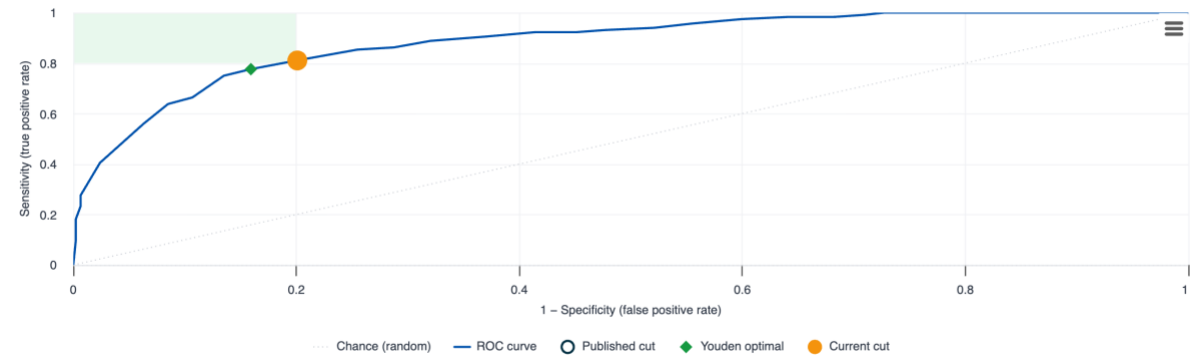
Decision Accuracies and Receiver Operating Curves for Fall Screenings

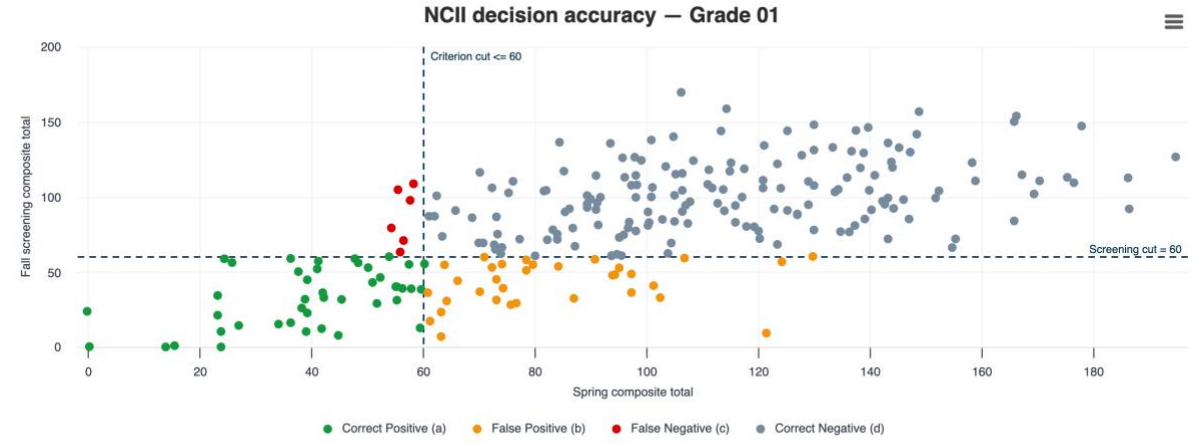


Open CP (a) students | Open FP (b) students | Open FN (c) students | Open CN (d) students

ROC curve — grade K

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.

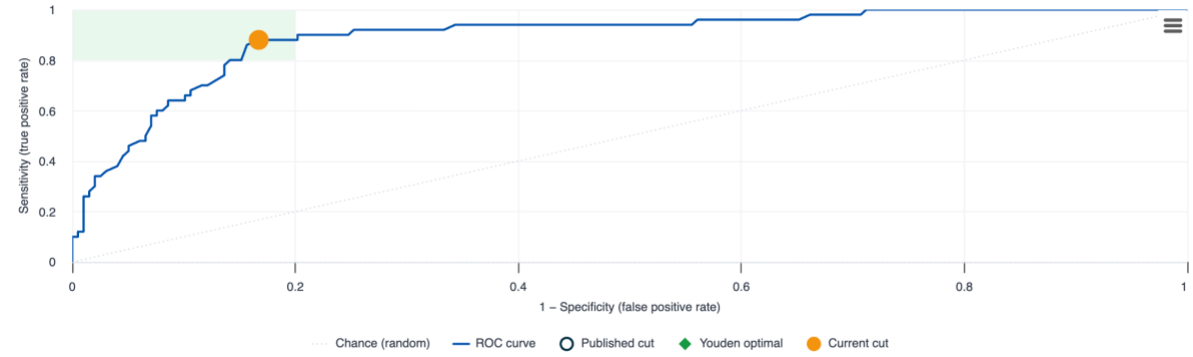


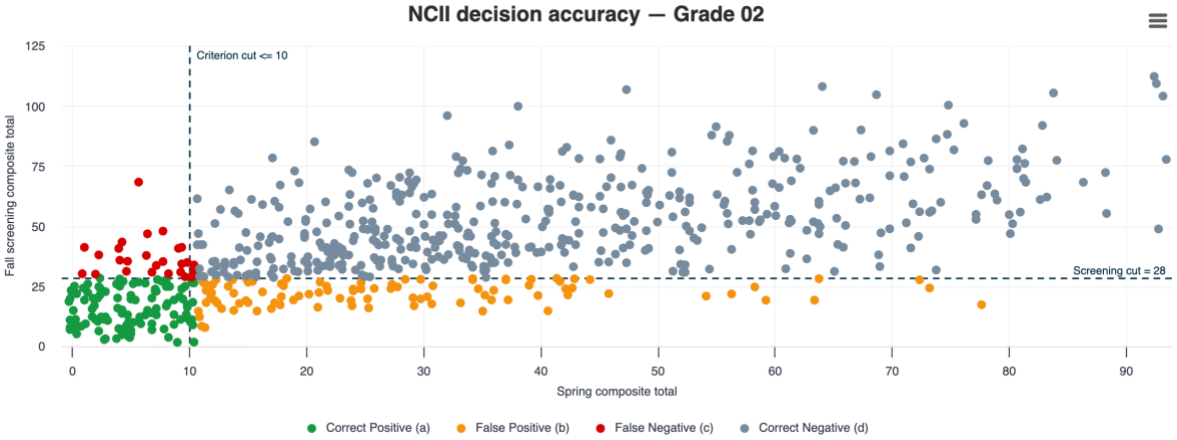


Open CP (a) students | Open FP (b) students | Open FN (c) students | Open CN (d) students

ROC curve — grade 01

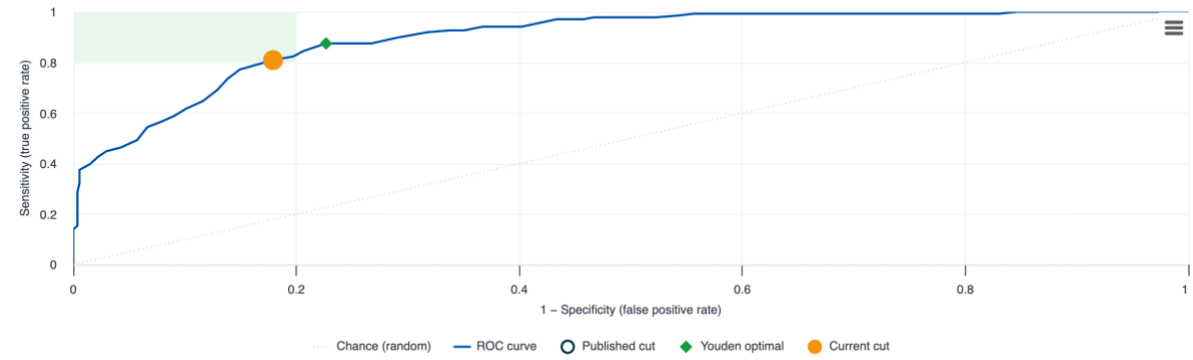
Sensitivity vs. 1 − specificity across every possible screener cut. The faint green band marks the NCII Full Bubble qualifying region (sens ≥ 0.80, FPR ≤ 0.20). Drag the open circle to test a hypothetical cut — the analysis-table row above will recompute its CP/FP/FN/CN, sens, spec, PPV, and NPV. AUC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

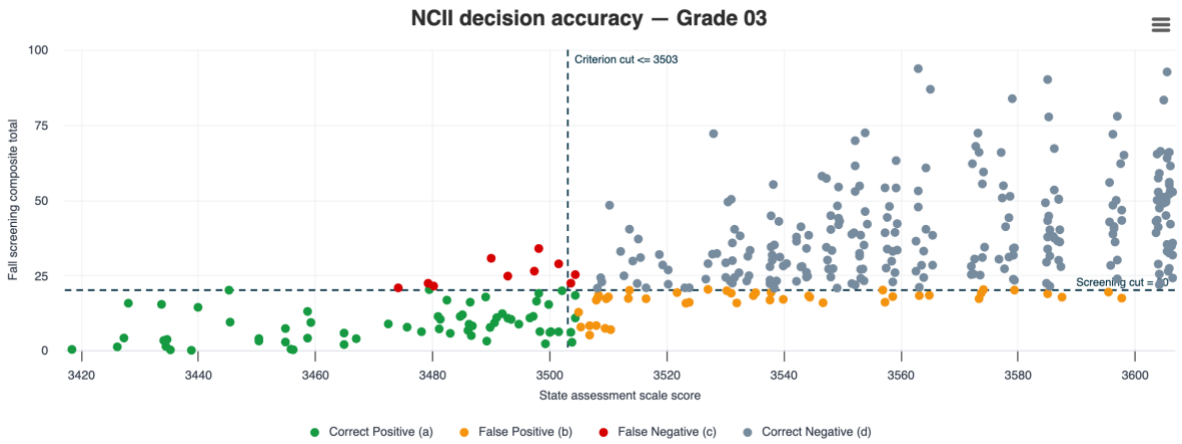




ROC curve — grade 02

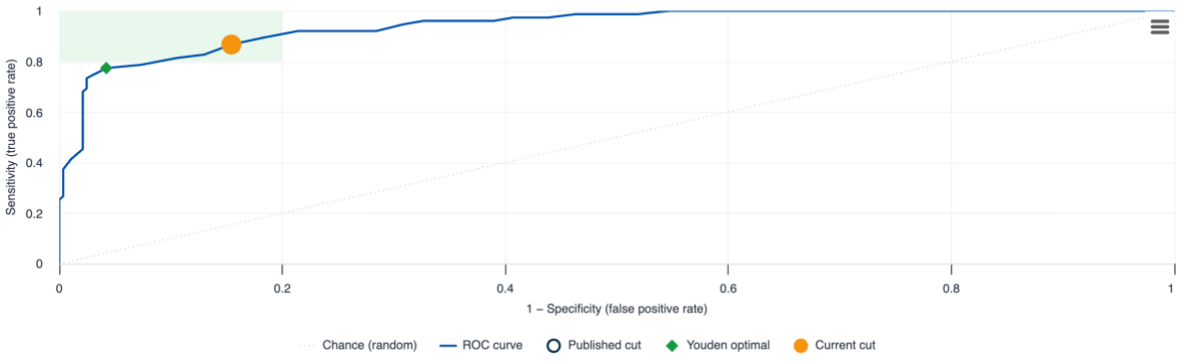
Sensitivity vs. 1 − specificity across every possible screener cut. The faint green band marks the NCII Full Bubble qualifying region (sens ≥ 0.80, FPR ≤ 0.20). Drag the open circle to test a hypothetical cut — the analysis-table row above will recompute its CP/FP/FN/CN, sens, spec, PPV, and NPV. AUC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

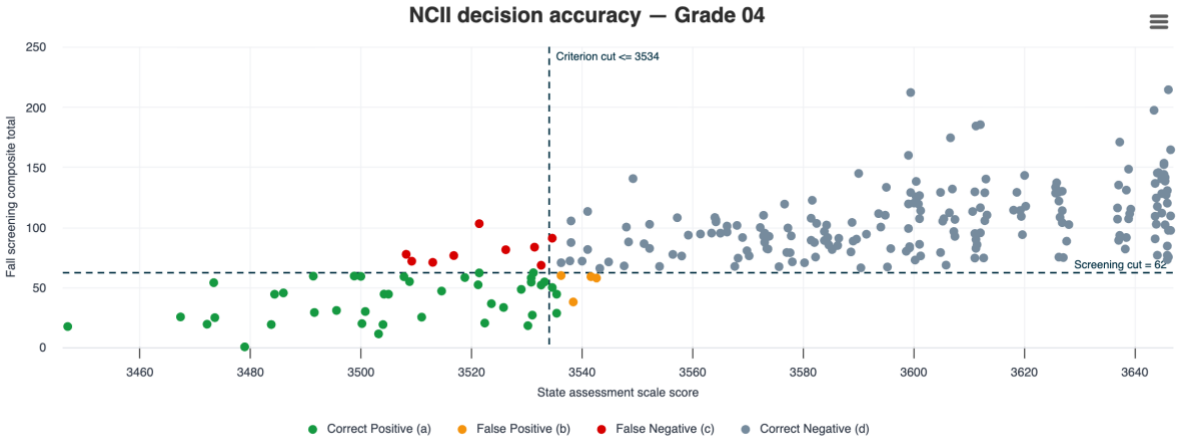




ROC curve — grade 03

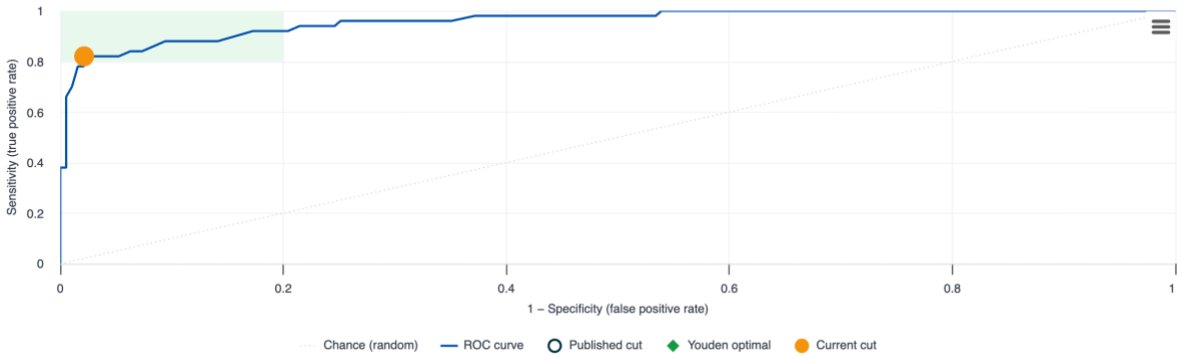
Sensitivity vs. 1 − specificity across every possible screener cut. The faint green band marks the NCII Full Bubble qualifying region (sens ≥ 0.80, FPR ≤ 0.20). Drag the open circle to test a hypothetical cut — the analysis-table row above will recompute its CP/FP/FN/CN, sens, spec, PPV, and NPV. AUC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

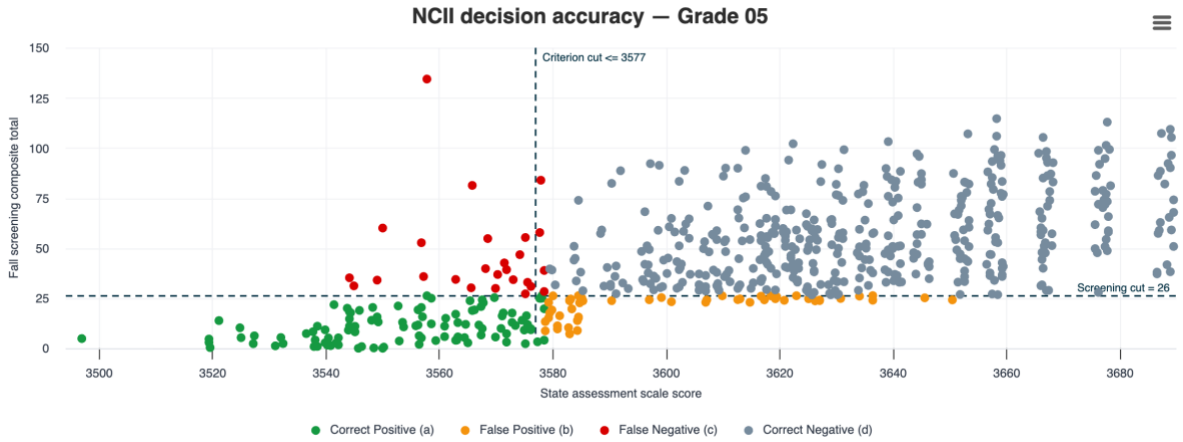




ROC curve — grade 04

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 05

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.