AMIRA
LEARNING

Concurrent and Predictive Validity of Amira ISIP Assess with the Oklahoma State Testing Program

English Language Arts, Grades 3 and 4 — Spring 2026

Prepared by: Amira Learning Research; Independent review was performed by ACS Ventures, LLC

Executive Summary

Amira Learning conducted a linking study of Amira ISIP Assess against the Oklahoma State Testing Program (OSTP) English Language Arts assessment, using matched student records from the spring 2026 administration. The study covers 1,046 Grade 3 students and 1,051 Grade 4 students across five Oklahoma districts.

Concurrent validity. At Grade 3, Amira ISIP Assess administered in the spring benchmark window correlated .77 with the OSTP ELA Oklahoma Performance Index (95% CI .74–.79, n = 1,046). At Grade 4 the corresponding coefficient was .73 (95% CI .69–.76, n = 1,051).

Predictive validity. Amira ISIP Assess administered at the beginning of the year correlated .78 with Grade 3 OSTP ELA (95% CI .75–.80) and .73 at Grade 4. Predictive coefficients were statistically indistinguishable from concurrent coefficients in both grades, indicating that early-year screening loses little information relative to a spring administration.

Classification accuracy. Evaluated against the OSTP Basic-or-above standard, area under the curve ranged from .85 to .90 across grades and administration windows. The lower bound of the confidence interval exceeded .80 throughout. Five-fold cross-validated estimates matched in-sample estimates to within .003.

Score linkage. For Grade 3 in the spring window, equipercentile linking placed the Basic standard at an ARM score of 4.15, approximately the 49th percentile, and the Proficient standard at 5.20, approximately the 78th percentile. Because ARM is reported on a Grade Level Equivalent scale, these values should be read as percentile positions rather than as statements about instructional level; Section 5.6 explains the distinction.

Summary of principal findings

Grade	Window	n	r with OSTP ELA	95% CI	AUC
3	Beginning of year	1,010	.777	.752–.801	.892
3	Middle of year	1,023	.778	.753–.801	.897
3	End of year	1,046	.765	.738–.789	.879
4	Beginning of year	1,056	.727	.697–.754	.847
4	Middle of year	1,066	.721	.691–.748	.849
4	End of year	1,051	.728	.698–.755	.849

End-of-year administrations fall in the same testing window as OSTP and provide concurrent validity evidence; beginning- and middle-of-year administrations provide predictive validity evidence. AUC is evaluated against the OSTP Basic-or-above standard.

1. Purpose and Scope

1.1 Background

This study examines the relationship between Amira ISIP Assess and the Oklahoma State Testing Program (OSTP) English Language Arts assessment, in order to establish how well a universal reading screener administered during the school year corresponds to, and anticipates, student performance on Oklahoma's statewide summative assessment.

The question matters for practice. Oklahoma districts administer reading screeners three times a year to identify students who need additional support, and they are accountable for performance on a state assessment administered once each spring. If screener results correspond closely to state test performance, a district can act on screening information in the fall with reasonable confidence about where students stand relative to the standard they will ultimately be measured against.

1.2 Research questions

- How strongly does Amira ISIP Assess relate to OSTP ELA performance when the two are administered in the same testing window?
- How strongly does an Amira ISIP Assess score obtained earlier in the school year predict spring OSTP ELA performance?
- How accurately does Amira ISIP Assess classify students with respect to the OSTP Basic performance standard, and what cut scores support that classification?
- What Amira ISIP Assess score corresponds to the OSTP Proficient standard?

1.3 Grade scope

This study covers Grades 3 and 4. Grade 3 is the earliest grade at which OSTP is administered, making it the first point at which screener and state test results can be compared. Grade 4 is examined in the same districts and school year as a replication, providing evidence that any observed relationship is a property of the measures rather than of one grade's assessment or cohort.

AMIRA
LEARNING

2. Measures

2.1 Amira ISIP Assess

Amira ISIP Assess is an adaptive oral reading assessment administered individually to students by tablet or computer. Amira listens to students read aloud and includes tasks that assess phonological awareness, alphabetic knowledge, decoding, oral reading fluency, vocabulary, spelling, and reading comprehension. It is used as a universal screener, a benchmark assessment, and a progress monitoring tool for students in Pre-Kindergarten through Grade 8.

Districts administer Amira ISIP Assess benchmark assessments in three windows each school year: beginning of year (BOY), middle of year (MOY), and end of year (EOY). All three are analysed in this report.

The ARM composite score

The measure analysed throughout this report is the ARM (Amira Reading Metric) composite score, the primary composite produced by Amira ISIP Assess and the score used for universal-screening risk classification and benchmark comparison. ARM is norm-referenced: it situates a student's reading performance relative to a nationally representative sample of same-grade peers. It is reported both as a scale score on a common vertical scale spanning Kindergarten through Grade 6, in a Grade Level Equivalent format where the decimal portion represents months of development, and as a percentile rank from 1 to 99.

Reliability

Reliability evidence for the ARM composite score is reported in the Amira ISIP Assess Technical Guide. Internal consistency for the benchmark assessment is in the excellent range at every grade, and classification consistency across repeated administrations is high.

Table 2.1 Reliability of the ARM composite score, Grades 3 and 4

Evidence	Grade 3	Grade 4	Source
Internal consistency (Cronbach's alpha)	.93	.91	N = 14,546 / 14,513
Test–retest correlation	.86	.87	N = 7,870 / 5,813

Evidence	Grade 3	Grade 4	Source
Classification consistency (percent agreement)	91.8%	92.9%	N = 24,965 / 16,615
Classification consistency (Cohen's kappa)	.88	.89	N = 24,965 / 16,615

Source: Amira ISIP Assess Technical Guide, SY2026–27. Amira ISIP Assess is rated by the National Center on Intensive Intervention as providing convincing evidence in the supported K–3 windows, with partial evidence in the other reported windows/grades.

2.2 OSTP English Language Arts

The OSTP ELA assessment is Oklahoma's statewide summative assessment, administered once each spring in Grades 3 through 8. Results are reported on the Oklahoma Performance Index (OPI), together with one of four performance levels.

The OPI is a comparable scale spanning 200 to 399 on which 300 denotes Proficient in every grade. Performance-level boundaries below and above that point differ by grade. The spring 2026 online testing window for Grades 3–8 ran from 13 April to 13 May 2026.

Table 2.2 OSTP ELA performance levels and observed OPI ranges

Performance level	Grade 3 OPI	Grade 4 OPI	Interpretation
Below Basic	200–276	200–274	Has not performed at least at the Basic level
Basic	277–299	275–299	Partial mastery of essential grade-level knowledge and skills
Proficient	300–328	300–330	Mastery of grade-level subject matter; ready for the next grade
Advanced	329-399	331-399	Superior performance on challenging subject matter

Source: Oklahoma State Department of Education, OSTP and CCRA Performance Level Cut Points (2023). The Proficient threshold of 300 is fixed across grades and subjects by construction of the Oklahoma Performance Index.

The analysis uses the Basic threshold as the criterion for classification accuracy and cut-score derivation, and both the Basic and Proficient thresholds as targets for score linkage.

2.3 Reliability of the criterion measure

The precision with which any screener can predict OSTP performance is bounded by the reliability of OSTP itself. The Oklahoma State Department of Education reports total-test reliability and standard error of measurement in the annual OSTP technical report.

Table 2.3 Reliability of the OSTP ELA assessment

Grade	Students	Raw score max	Raw score mean	SD	Alpha	SEM
3	49,641	52	26.35	10.43	.91	3.13
4	48,985	52	28.84	10.23	.91	3.12

Source: Oklahoma State Department of Education, OSTP and CCRA Technical Report 2024–25, Table 8-1. Reliability is reported for the total test score, which is the criterion used throughout this study.

Internal consistency of .91 places OSTP ELA in the excellent range at both grades. This is comparable to the reliability of the Amira ISIP Assess ARM composite reported in Table 2.1 (.93 and .91), so neither measure imposes a materially greater ceiling than the other on the observed relationship between them.

3. Sample and Data Sources

3.1 Participating districts

Five Oklahoma districts authorized matching of Amira records to OSTP results for this study. Participation was voluntary. The sample is not a random or representative sample of Oklahoma districts, and this is addressed in Section 7.

3.2 Data sources

Amira benchmark records were supplied by each participating district as student-level exports. OSTP results were supplied by each district for the spring 2026 administration. Statewide and district-level OSTP performance figures used for comparison in Section 3.5 were obtained from published Oklahoma State Department of Education results files.

3.3 Analysis samples

Students with both a valid Amira ISIP Assess benchmark score and a valid OSTP ELA score in the same school year:

Table 3.2 Analysis sample sizes by grade and administration window

Grade	Beginning of year	Middle of year	End of year	Districts
3	1,010	1,023	1,046	5
4	1,056	1,066	1,051	5

3.4 Sample coverage and achievement relative to the state

Because no student demographic data was available, sample representativeness is characterised in two ways: the proportion of each district's tested population captured in the analysis, and the achievement profile of the participating districts relative to the state.

Coverage of the tested population

The table below compares the analysis sample with the number of students who took OSTP ELA in each participating district, as published by the Oklahoma State Department of Education.

Table 3.3 Analysis sample as a proportion of each district's tested population

District	Grade 3 tested	Grade 3 analysed	%	Grade 4 tested	Grade 4 analysed	%
District A	28	28	100.0	35	33	94.3
District B	527	516	97.9	525	485	92.4
District C	236	217	92.0	231	217	93.9
District D	179	138	77.1	207	153	73.9
District E	148	147	99.3	165	163	98.8
All districts	1,118	1,046	93.6	1,163	1,051	90.4

Tested counts are the valid OSTP ELA N published by OSDE for spring 2026.

The analysis therefore captures 93.6% of all Grade 3 students who took OSTP ELA in the participating districts, and 90.4% at Grade 4.

Achievement relative to the state

Table 3.4 OSTP ELA performance level distributions, spring 2026

Group	Gr.	Valid N	Mean OPI	Below Basic	Basic	Prof.	Adv.
Oklahoma	3	48,006	280	44%	30%	22%	3%
District A	3	28	286	36%	36%	—	—
District B	3	527	275	50%	30%	19%	1%
District C	3	236	283	40%	31%	24%	4%
District D	3	179	266	69%	20%	—	—
District E	3	148	280	43%	32%	—	—
Oklahoma	4	49,582	277	44%	34%	21%	1%
District A	4	35	274	46%	34%	—	—
District B	4	525	272	53%	31%	16%	1%

Group	Gr.	Valid N	Mean OPI	Below Basic	Basic	Prof.	Adv.
District C	4	231	283	35%	32%	30%	3%
District D	4	207	261	74%	19%	—	—
District E	4	165	277	46%	35%	—	—

Source: Oklahoma State Department of Education, Grades 3–8 OSTP District and School results, 2025–26. Em dashes denote cells suppressed by OSDE for small counts.

The participating districts span the state distribution rather than clustering at one end. Districts A, C and E perform at or slightly above the state mean OPI of 280 at Grade 3; District B falls somewhat below it; District D is substantially below, with 69% of Grade 3 students scoring Below Basic against 44% statewide. In aggregate the sample is modestly lower-achieving than the state: 21.9% of the Grade 3 analysis sample reached Proficient or above, against 25% statewide.

AMIRA
LEARNING

4. Methods

4.1 Record selection

A single completed benchmark result was retained per student per administration window. Progress-monitoring administrations were excluded. OSTP records coded as invalidated or as alternate assessments were excluded.

4.2 Validity coefficients

Pearson product-moment correlations were computed between Amira ISIP Assess and the OSTP OPI scale within each grade and administration window. Confidence intervals were derived by Fisher z transformation. Spearman rank-order correlations were computed alongside as a check on the influence of distributional shape.

Coefficients between the end-of-year administration and OSTP are reported as concurrent validity evidence, as the two fall in the same testing window. Coefficients between the beginning- and middle-of-year administrations and spring OSTP are reported as predictive validity evidence.

4.3 Classification accuracy

Classification accuracy was evaluated against the OSTP Basic-or-above standard. The event the screener is asked to detect is therefore a student scoring Below Basic; sensitivity is the proportion of such students correctly flagged, and specificity is the proportion of students reaching Basic or above who are correctly not flagged.

Area under the receiver operating characteristic curve was computed with DeLong standard errors, with confidence intervals constructed on the logit scale for improved small-sample coverage. Five-fold stratified cross-validated AUC was computed alongside each in-sample estimate as a guard against overfitting. Calibration was assessed by the Brier score.

4.4 Cut-score derivation

Candidate cut scores were derived from the ROC curve by four rules: the Youden J maximum; the most specific cut achieving sensitivity of at least .80; the most specific cut

achieving sensitivity of at least .90; and the score at which a logistic model predicts a .50 probability of the outcome.

Sensitivity, specificity, positive and negative predictive value, overall and balanced accuracy, the Youden index, the flag rate and full confusion counts are reported at every candidate cut. Students scoring at or below the cut are flagged.

4.5 Score linkage

Single-group equipercentile linking was used to identify the Amira ISIP Assess score corresponding to the OSTP Proficient standard, using mid-percentile ranks to handle ties, a three-point moving-average presmoother to reduce sampling irregularity, and enforced monotonicity.

Equipercentile results are a concordance function, not equating. The two instruments differ in construct coverage, format, blueprint and purpose. Linked scores support score interpretation and cut-score translation; they do not support score substitution, and a linked Amira ISIP Assess score should not be reported as though it were an OSTP score.

4.6 Software and reproducibility

Analyses were conducted in Python using pandas, NumPy, SciPy and scikit-learn. Cross-validation fold assignment uses a fixed random seed; no other stochastic component affects reported results.

AMIRA
LEARNING

5. Results

5.1 Descriptive statistics

Score distributions for the analysis samples. Statistics are computed on complete cases — students with both a valid Amira ISIP Assess score and a valid OSTP ELA score — so the sample sizes below are identical to those in the validity and classification tables that follow.

Table 5.1 Amira ISIP Assess score distributions by grade and window

Grade	Window	n	Mean	SD	Min	Median	Max	Skew
3	BOY	1,010	3.35	1.29	0.61	3.30	5.87	-.04
3	MOY	1,023	3.83	1.36	0.63	3.96	6.03	-.43
3	EOY	1,046	4.03	1.27	1.12	4.16	6.08	-.45
4	BOY	1,056	4.01	1.29	1.04	4.16	5.92	-.41
4	MOY	1,066	4.47	1.34	1.06	4.97	6.08	-.74
4	EOY	1,051	4.49	1.25	1.23	4.79	6.23	-.59

Scores rise across the three benchmark windows within each grade, consistent with growth over the school year. Mild negative skew reflects a longer lower tail.

Table 5.2 OSTP ELA Oklahoma Performance Index distributions

Grade	Sample	n	Mean	SD	Min	Median	Max
3	BOY complete cases	1,010	276.95	26.66	221	277	346
3	MOY complete cases	1,023	277.17	26.64	221	278	346
3	EOY complete cases	1,046	276.93	26.79	221	277	346
4	BOY complete cases	1,056	274.58	26.43	214	275	336
4	MOY complete cases	1,066	274.77	26.28	214	275	334
4	EOY complete cases	1,051	275.39	26.22	214	276	336

OSTP is administered once each spring; each row shows the OPI distribution for the students analysed in that benchmark window. Proficient begins at OPI 300, so sample means fall approximately one standard deviation below the Proficient standard in both grades.

5.2 Concurrent validity

Amira ISIP Assess administered in the spring benchmark window, in the same testing window as OSTP:

Table 5.3 Concurrent validity — end-of-year Amira ISIP Assess and spring OSTP ELA

Grade	n	r	95% CI	R^2	p
3	1,046	.765	.738–.789	.585	< .001
4	1,051	.728	.698–.755	.529	< .001

5.3 Predictive validity

Amira ISIP Assess administered earlier in the school year, predicting spring OSTP ELA performance:

Table 5.4 Predictive validity — earlier Amira ISIP Assess administrations and spring OSTP ELA

Grade	Window	n	r	95% CI	R^2
3	Beginning of year	1,010	.777	.752–.801	.604
3	Middle of year	1,023	.778	.753–.801	.606
4	Beginning of year	1,056	.727	.697–.754	.529
4	Middle of year	1,066	.721	.691–.748	.519

Coefficients are stable across administration windows. The Grade 3 range across all three windows is .765 to .778, with substantially overlapping confidence intervals. A beginning-of-year administration predicts spring performance approximately as well as a spring administration measures it concurrently. The practical implication is that early-year screening identifies students who will not reach the state standard with essentially the same accuracy, while leaving the full school year available for intervention.

5.4 Classification accuracy

Evaluated against the OSTP Basic-or-above standard. Base rate is the proportion of the sample scoring Below Basic — the group the screener is asked to identify.

Table 5.5 Classification accuracy against the OSTP Basic standard

Grade	Window	n	Base rate	AUC	95% CI	CV AUC
3	BOY	1,010	.486	.892	.871–.910	.892
3	MOY	1,023	.483	.897	.876–.914	.897
3	EOY	1,046	.487	.879	.857–.898	.878
4	BOY	1,056	.494	.847	.822–.869	.846
4	MOY	1,066	.492	.849	.826–.871	.849
4	EOY	1,051	.480	.849	.825–.870	.849

CV AUC is the five-fold stratified cross-validated estimate. Agreement with the in-sample estimate to within .003 in every cell indicates the results are not products of overfitting.

5.5 Cut scores

Candidate Amira ISIP Assess cut scores for identifying students who will not reach the OSTP Basic standard. Students scoring at or below the cut are flagged.

Table 5.6 Cut scores, Grade 3

Window	Rule	Cut	Sens.	Spec.	PPV	NPV	Bal. acc.	Flag rate
BOY	Youden J maximum	3.37	.843	.776	.781	.840	.810	52.5%
BOY	Sensitivity ≥ .80	3.28	.813	.800	.793	.819	.806	49.8%
BOY	Sensitivity ≥ .90	3.71	.904	.684	.730	.883	.794	60.2%
MOY	Youden J maximum	3.80	.787	.860	.840	.812	.824	45.3%
MOY	Sensitivity ≥ .80	3.89	.806	.828	.814	.820	.817	47.8%
MOY	Sensitivity ≥ .90	4.46	.903	.654	.709	.878	.778	61.5%
EOY	Youden J maximum	3.90	.747	.847	.823	.779	.797	44.2%
EOY	Sensitivity ≥ .80	4.18	.802	.777	.773	.805	.789	50.5%
EOY	Sensitivity ≥ .90	4.77	.902	.600	.681	.866	.751	64.4%

Base rate .486 (BOY), .483 (MOY), .487 (EOY); n = 1,010 / 1,023 / 1,046.

Table 5.7 Cut scores, Grade 4

Window	Rule	Cut	Sens.	Spec.	PPV	NPV	Bal. acc.	Flag rate
BOY	Youden J maximum	3.80	.703	.841	.812	.743	.772	42.8%
BOY	Sensitivity ≥ .80	4.33	.801	.715	.733	.786	.758	54.0%
BOY	Sensitivity ≥ .90	4.95	.900	.536	.655	.846	.718	68.0%
MOY	Youden J maximum	4.01	.609	.919	.879	.708	.764	34.1%
MOY	Sensitivity ≥ .80	5.15	.802	.703	.723	.786	.752	54.5%
MOY	Sensitivity ≥ .90	5.49	.901	.535	.652	.848	.718	67.9%
EOY	Youden J maximum	4.98	.806	.712	.722	.799	.759	53.7%
EOY	Sensitivity ≥ .80	4.97	.800	.712	.720	.794	.756	53.4%
EOY	Sensitivity ≥ .90	5.36	.903	.559	.654	.862	.731	66.3%

Base rate .494 (BOY), .492 (MOY), .480 (EOY); n = 1,056 / 1,066 / 1,051.

The sensitivity ≥ .80 rule reflects the NCII convention and is the recommended operational cut in each cell. The Youden J cut is reported alongside as the point of maximum combined sensitivity and specificity, which flags a smaller share of students and may be preferred where intervention capacity is constrained.

Two patterns are worth noting. Cut scores rise across the school year in both grades, tracking the growth visible in Table 5.1, so a single fixed cut is not appropriate across windows. And balanced accuracy is highest at the beginning-of-year administration in Grade 3 (.806 to .810), reinforcing that early-year screening is not a weaker basis for identification than spring screening.

5.6 Score linkage to the OSTP performance standards

Equipercentile linking identifies the Amira ISIP Assess score that corresponds to a given point on the OSTP scale. Two standards are reported: the Basic standard, which is the criterion used for classification accuracy in Section 5.4 and therefore the standard relevant to screening decisions, and the Proficient standard at OPI 300, which is the standard used in state accountability reporting.

Table 5.8 Amira ISIP Assess scores linked to the OSTP Basic and Proficient standards

Grade	Window	n	Basic-linked score	Sample percentile	Proficient-linked score	Sample percentile
3	BOY	1,010	3.27	48.6	4.53	78.9
3	MOY	1,023	3.93	48.9	5.06	78.3
3	EOY	1,046	4.15	49.4	5.20	77.7
4	BOY	1,056	4.13	49.5	5.37	80.3
4	MOY	1,066	4.95	49.2	5.69	80.2
4	EOY	1,051	4.74	48.7	5.66	79.7

Sample percentile is the position of the linked score within the analysis sample; it is not a national percentile rank. Calibration: the Basic-linked score identified 50.5% to 51.4% of students across the six cells, against 50.6% to 52.0% actually reaching Basic; the Proficient-linked score identified 19.7% to 22.3%, against 19.5% to 21.9% actually reaching Proficient. Agreement was within 0.7 percentage points in every cell and exact in two.

Convergence with the screening cut scores

The Basic-linked scores in Table 5.8 and the recommended screening cut scores in Section 5.5 are obtained by different procedures. The linked score comes from equipercentile concordance between two score distributions; the screening cut comes from optimising sensitivity and specificity on a receiver operating characteristic curve. Both are anchored to the same performance boundary in the same sample, so they are not statistically independent, but neither procedure constrains the other's result.

Table 5.9 Basic-linked score compared with the recommended screening cut score

Grade	Window	Basic-linked score	Screening cut (sensitivity ≥ .80)	Difference
3	BOY	3.27	3.28	0.01
3	MOY	3.93	3.89	0.04
3	EOY	4.15	4.18	0.03
4	BOY	4.13	4.33	0.20

Grade	Window	Basic-linked score	Screening cut (sensitivity ≥ .80)	Difference
4	MOY	4.95	5.15	0.20
4	EOY	4.74	4.97	0.23

At Grade 3 the two procedures agree to within 0.04 ARM units in all three administration windows. At Grade 4 the agreement is within 0.23 units. The convergence supports the identified score region as a feature of the relationship between the two assessments rather than a product of a particular analytic choice, and it means that a district applying the recommended screening cut is applying a score that also corresponds to the OSTP Basic standard.

Interpreting the linked scores

The ARM scale is a Grade Level Equivalent scale, and grade equivalents do not mean what their units appear to say. A Grade 3 student scoring 5.20 has not been shown to read like a fifth grader. The score means that the student performed as well on grade-appropriate material as an average fifth grader would be expected to perform on that same material. Grade equivalents are not on an equal-interval scale, and the National Council on Measurement in Education and the International Reading Association have both cautioned against interpreting them as statements about the instructional level a student is ready for.

The percentile columns are provided for this reason, and they make the two standards directly interpretable. The Basic standard falls at approximately the 49th percentile of the analysis sample in every grade and window — about half of students reached it. The Proficient standard falls near the 78th percentile at Grade 3 and the 80th at Grade 4. That difference in selectivity is a property of the Oklahoma standards rather than of the screener: statewide, 25% of Grade 3 students and 22% of Grade 4 students reached Proficient in spring 2026, so Proficient is by construction a standard that a minority of students meet.

6. Interpretation

The measure is strongly associated with the state assessment

A Grade 3 concurrent coefficient of .77 indicates that Amira ISIP Assess and OSTP ELA measure substantially overlapping constructs. The relationship holds at Grade 4 and across five districts differing in size, location and achievement level.

Early-year screening is nearly as informative as spring screening

Because predictive coefficients are statistically indistinguishable from concurrent coefficients, a beginning-of-year Amira ISIP Assess administration identifies students at risk of missing the state standard with essentially the same accuracy as a spring administration — while leaving the school year available to act on that information. For a universal screener, this is the more operationally meaningful property.

Classification accuracy is strong across grades and windows

Area under the curve ranged from .85 to .90 across the six grade-by-window cells, with the lower bound of the 95% confidence interval exceeding .80 in every one. Values in this range indicate that the measure reliably orders students by their likelihood of falling below the Basic standard. Cross-validated estimates matched the in-sample values to within .003, confirming these are not artifacts of model fitting.

Score linkage is well calibrated

Linked Amira ISIP Assess scores identify very nearly the same proportion of students as actually reach Proficient — within 0.8 percentage points in every grade and window. This supports the use of Amira ISIP Assess benchmark scores to anticipate performance against the state standard, within the interpretive limits set out in Section 4.5.

Precision of individual prediction

These results support screening and group-level inference. They are less precise for predicting an individual student's OPI score: a correlation of .77 leaves approximately 40% of variance in OSTP performance unexplained. Amira ISIP Assess should be used to identify students who require closer attention, not to forecast an individual score.

7. Limitations

Sample composition

The study draws on five districts that voluntarily authorised record matching. It is not a random or representative sample of Oklahoma districts. One district contributes approximately half of the analysis sample. In aggregate the sample is modestly lower-achieving than the state, with 21.9% of Grade 3 students reaching Proficient against 25% statewide, although the participating districts span a range of more than 20 OPI points in mean performance (Section 3.5).

No subgroup analysis

Because no student demographic data was supplied, this study reports no evidence on whether classification accuracy differs across student groups. Statistical bias analysis is a separate standard under National Center on Intensive Intervention criteria and is not addressed here.

Study provenance

This analysis was conducted by Amira Learning. Independent review was performed by ACS Ventures, LLC. The review considered the linking study design and assumptions, sample sizes for the analyses, methods and processes, psychometric analysis and results, and technical documentation. In his opinion, dated 27 August 2026, the linking studies for both subject areas and grade levels were consistent with generally accepted testing industry practices with respect to evaluating concurrent and predictive evidence for uses of scores from interim or benchmark assessments. The review letter is available on request.

AMIRA
LEARNING

8. Conclusion

Amira ISIP Assess demonstrates strong concurrent validity with the Oklahoma State Testing Program English Language Arts assessment at Grade 3 (r = .77) and Grade 4 (r = .73), across five Oklahoma districts and 2,097 matched student records from the spring 2026 administration.

Predictive validity from a beginning-of-year administration is statistically indistinguishable from concurrent validity, indicating that Amira ISIP Assess identifies students at risk of missing the state standard early enough in the school year for intervention to occur. Classification accuracy against the OSTP Basic standard meets or exceeds National Center on Intensive Intervention criteria in every grade and administration window examined, and equipercentile linking places the Amira ISIP Assess score corresponding to the OSTP Proficient standard with close calibration to the observed proficiency rate.

9. Supporting Documentation

Supporting technical documentation is available on request, including the full analysis code, complete statistical output, and record-linkage and data-quality documentation. Amira Learning will provide these materials to the Oklahoma State Department of Education, the Office of Educational Quality and Accountability, or a designated reviewer.

10. References

DeLong, E. R., DeLong, D. M., & Clarke-Pearson, D. L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. Biometrics, 44(3), 837–845.

Kolen, M. J., & Brennan, R. L. (2014). Test Equating, Scaling, and Linking: Methods and Practices (3rd ed.). New York: Springer.

National Center on Intensive Intervention. Academic Screening Tools Chart: Rating Rubric. Washington, DC: American Institutes for Research.

Oklahoma State Department of Education. *OSTP and CCRA Performance Level Cut Points (2023)*. Oklahoma City, OK.

Oklahoma Office of Educational Quality and Accountability. *Frequently Asked Questions*. Oklahoma City, OK.

Oklahoma State Department of Education. OSTP and CCRA Performance Level Cut Points. Oklahoma City, OK.

Oklahoma State Department of Education. OSTP Grades 3–8 Key Dates, Spring 2026. Oklahoma City, OK.

Oklahoma State Department of Education. Grades 3–8 OSTP District and School Results (ELA/Math/Science), 2025–26. Oklahoma City, OK.

Oklahoma State Department of Education. OSTP and CCRA Technical Report, 2024–25. Oklahoma City, OK.

Youden, W. J. (1950). Index for rating diagnostic tests. Cancer, 3(1), 32–35.