AMIRA LEARNING

White Paper

Delivering Equity:

Evidence that Amira's Reading Assessment Tests Every Learner Fairly and Accurately

May, 2026

A consolidated review of training methodology, psychometric students' subgroup performance, year-over-year outcomes for English learners, and independent district-level scoring validation studies.

Istation.com + AmiraLearning.com | Amira-Elevated Learning

AMIRA
LEARNING

Executive Summary

This paper reviews the evidence that Amira scores students with accents and dialects fairly and accurately, with a particular focus on New Mexico.

Amira is the most widely used reading screener and tutoring system in the United States. It is used across districts and on state-approved adoption lists in nearly every region of the country. Because Amira listens to children read aloud, it is essential to ask whether Amira listens equitably for students of color, English learners, and students who speak with regional, ethnic, or first-language-influenced accents and dialects.

The evidence assembled in this paper supports a clear conclusion. Amira is designed for equity, trained for equity, validated for equity, and empirically demonstrates equity. The data comes from internal studies, third-party state agency reviews, district-level validation studies, and year-over-year outcome analyses in real classrooms.

Five lines of evidence

1. **Trained For Equity:** Amira is designed from the ground up to detect reading skill gaps, not accent or dialect differences. Its training corpus, annotation protocol, and model architecture were all built around this distinction.
2. **Psychometrically Equitable:** Amira's reliability and item-level fairness hold consistently across every measured student group, with reliability scores at or above 0.90 for nearly every grade and student subgroup. (A reliability score of 0.80 or higher is considered strong; 0.90 is excellent).
3. **Comparatively Equitable:** In New Mexico, where the question of voice scoring and English learner performance has been directly tested, English learner (EL) proficiency parity with non-English Learners (non-EL) improved by 36% year-over-year after the introduction of Amira.
4. **Human Validated Equitable:** Independent district-level scoring validation studies — most recently a Spring 2026 study in Texas ISD spanning 1,827 oral reading fluency activities and over 301,000 words scored — confirm 96.9% overall human-Amira agreement, with 95.8% accuracy for non-English native language speakers.
5. **Cause of Misperception:** Where scoring issues do occur, the root cause is consistently traceable to audio environment, rostering, or protocol fidelity — not to the Amira's underlying accuracy. When implementation conditions are met, Amira's student subgroup accuracy is at least as high as that of the overall population.

AMIRA LEARNING

The bottom line

Amira does not score reading performance based on whether a student pronounces to a standard of articulation. Amira's scores are based on whether the student has accurately decoded the phonemes of the target text. The data in this paper demonstrate that this design philosophy translates into measurable equity outcomes across all populations studied.

The Evidence in Plain Terms

This section summarizes the evidence supporting Amira's equity. Each claim is documented in the full paper that follows.

If Amira were unable to listen accurately to ELLs, then Amira would be giving those ELLs LOW scores, failing to credit them with the reading mastery they actually possess. The data shows, in fact, that this year in New Mexico, Amira is flagging 26% FEWER ELLs than the keyboard test in previous years. This shows that Amira is doing a better job of driving out bias than previous New Mexico assessments.

If Amira were unable to listen accurately to ELLs, then Amira would show less psychometric accuracy for those students than other assessments. In fact, in submissions to State Education Agencies, Amira's psychometric reliability for ELLs is better than for leading assessments.

If Amira were unable to accurately listen to black students and other groups whose pronunciation follows distinct patterns, then Amira would show lower psychometric accuracy for those students than for other assessments. In fact, in submissions to State Education Agencies, Amira's psychometric reliability for black students is better than for leading assessments.

If Amira were unable to listen accurately to non-white students, then Amira would show lower inter-rater reliability on those students than on white students. In fact, inter-rater testing shows that Amira is MORE accurate for non-white students.

If Amira were biased, it would show larger score differences between white students and non-white student groups than New Mexico's previous assessments did. In fact, Amira shows smaller differentials, demonstrating that it is overcoming the bias of standard testing.

If Amira were biased, statistical tests like DIF would do what they are designed to do and reveal that bias. In fact, Amira has pristine DIF scores, demonstrating that Amira's items are free from bias.

If Amira struggled to listen to some students but not others, teachers who re-scored Amira would make more corrections on ELLs and other student groups than on white students. In fact, field testing shows that correction levels are statistically equivalent.

Science is about formulating a hypothesis and testing it experimentally and empirically, rather than relying on anecdotes and isolated incidents. The hypothesis that Amira has a sub-population-based bias is disproven by 7 separate definitive analyses.

1. Amira is designed for equity

Equity in Amira-powered assessment is not an outcome that can be added at the end. Equity is a property of the system's design, training, and testing. Amira treats equity as the core technical constraint shaping every aspect of the Assessment.

Designed For Equity

Amira does not "recognize speech" or attempt to turn speech into text. This avoids the central challenge of accent and dialect.

Most discussions of Amira's fairness in speech-enabled education assume that the underlying technology is automated speech recognition (ASR) — a system designed to transcribe what a speaker is saying. Amira is not ASR. Amira is a Reading Error Detection (RED) system.

The difference is fundamental. An ASR system is grounded in what the speaker says. Amira is grounded in what the student is supposed to read. Because Amira already knows the target text, its job is not to transcribe open-ended speech — it is to determine whether the student has accurately decoded the phonemes of the words in front of them.

This distinction has direct equity implications. When commercial ASR systems exhibit elevated word error rates for speakers of African American English or those with non-native accents, the gap is typically driven by the ASR's attempt to convert non-standard speech into standard transcription. Amira deliberately avoids this problem. Amira is not trying to detect every speech deviation a student produces — it is trying to detect deviations that indicate reading skill gaps.

Design principle

Amira does not judge the purity of pronunciation, require an "American" pronunciation, nor completely discount accent and dialect when those features are indistinguishable from a reading error. Amira evaluates whether a student is reading the phonemes in an English word in a manner consistent with a teacher-normed expectation for sounding them out.

Trained for Equity

Amira is trained disproportionately on Latino, Black, and ELL student reading. Amira overweights these student groups to ensure Amira is not biased.

The Amira RED model is trained on a corpus of more than 100,000 hours of children's speech captured during real assessment sessions in real classrooms. This training data is not a curated laboratory dataset. Amira has learned from millions of students with diverse accents, dialects, ambient classroom conditions, and developmental stages encountered in production.

The training methodology was specifically designed to ensure that students from historically marginalized speech communities are not just represented but proportionally overrepresented relative to their share of the general student population:

- Half of the speech-layer training corpus comes from a large random sample of students enrolled in schools where students of color constitute more than 75% of enrollment, based on data from the National Center for Education Statistics (NCES).
- The other half comes from a representative sample of Amira's remaining usage data, drawn from districts nationwide.
- The expert-hand-annotated dataset used to train Amira and validate the end-to-end RED model spans thousands of students, with student subgroup composition deliberately weighted toward dialect- and accent-bearing populations: 25% African American, 28% U.S. multilingual learners (MLLs/ELLs), 33% Other U.S. students, and 14% international students.

This intentional overweighting of accent- and dialect-speaking populations ensures the model is exposed to sufficient examples to perform adequately on their speech — a design choice that distinguishes Amira from general-purpose ASR systems whose training data over-represents standard American English.

Training data only teaches a model of what people teach. Amira's annotators are explicitly trained to flag genuine reading errors in accordance with the science of reading, and — critically — to not label deviations attributable to accent or dialect as errors. If a student consistently substitutes one phoneme for another in a way that reflects a stable speech pattern — such as */w/* for */r/* in younger readers, or */h/* for */j/* for native Spanish-speaking English learners — these substitutions are not annotated as reading miscues. The annotators look for patterns that indicate decoding failure, not patterns that indicate accent. Because supervised learning models perform the way their training data is labeled, this annotation discipline is the single most important equity safeguard in the system.

Proofed for Equity

Amira's scoring decisions are continuously tested using a method called Differential Item Functioning (DIF) to detect and remove bias. DIF analysis is the industry-standard method for assessing whether a test question is harder for one group of students than for another, even when students have the same overall reading ability.

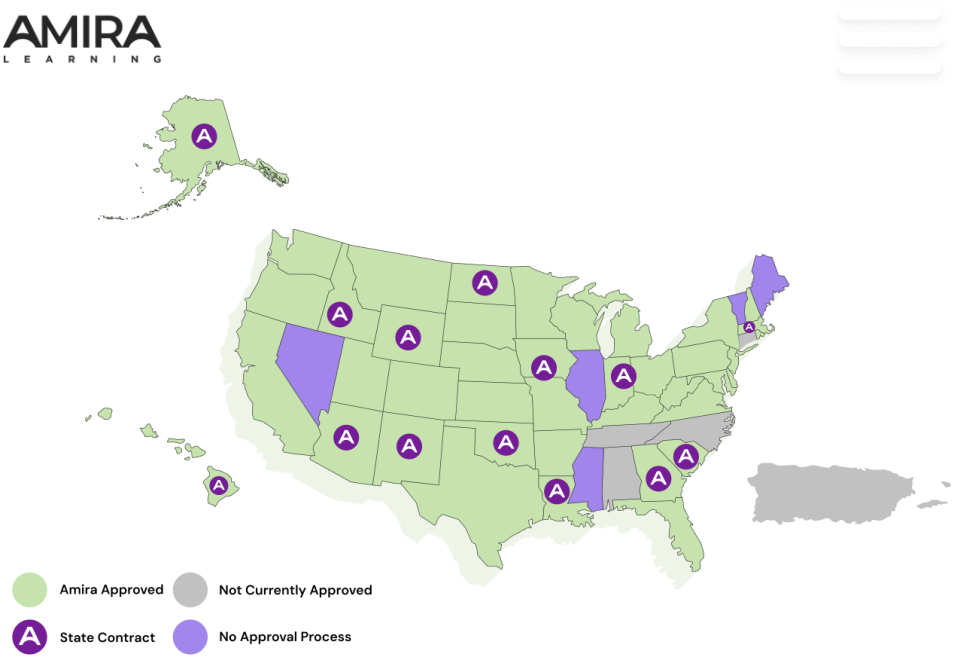
Amira's commitment to equity does not end at model deployment. Every iteration of the RED model is evaluated by disaggregating performance metrics across demographic and linguistic student subgroups, including race and ethnicity, English learner status, age and grade level, and overall reading ability. Any signs of degradation in any student subgroup trigger a detailed error analysis and adjustments to the model training process.

In parallel, the Amira ISIP Assessment item bank is routinely subjected to Differential Item Functioning (DIF) analysis using Item Response Theory (IRT) methods. DIF analysis identifies items in which performance differs across student subgroups at the same level of underlying ability — a statistical signature of item-level bias. Items that show DIF are reviewed annually by Amira's psychometric and instructional design teams and either recalibrated or removed from the active item bank.

Test-level DIF rates remain uniformly at or below 5% across every comparison studied (gender, Black/White, Hispanic/White), indicating an end-to-end equity result. A reasoned discussion about equity begins with real data, not unsubstantiated claims. The DIF results provide strong statistical evidence that Amira assesses students equitably, as DIF is the standard method for evaluating fairness across assessments.

Government-Validated For Equity

Amira has been evaluated by the psychometric teams employed by state education agencies in virtually every state that conducts such reviews. These reviews include an independent examination of correlation, reliability, validity, interrater reliability, and evidence of bias — the same evidence summarized in this paper. Amira is one of the most widely state-approved reading screeners in existence, and these approvals are not granted without scrutiny of the evidence presented here.



In an independent review conducted by the Sandra Dunagan Deal Center for Early Language and Literacy for the State of Georgia (November 2024), Amira ISIP ranked first among all screeners reviewed, with an overall score of 142.6 out of 155 — significantly above any other tool. Amira scored highest of all screeners on Technical Adequacy, earning 19.6 out of a possible 20 points. It was subsequently selected as the sole screener provided free of charge to all Georgia school districts.

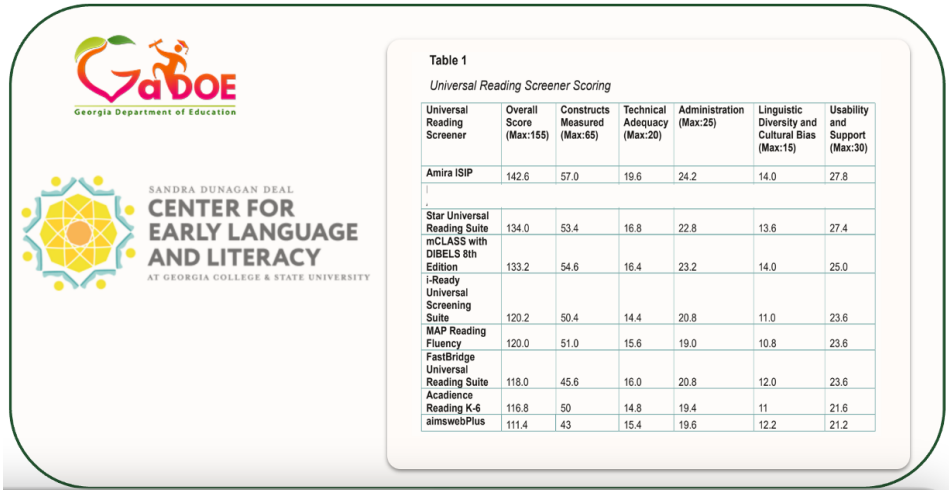


Table 1

Universal Reading Screener Scoring

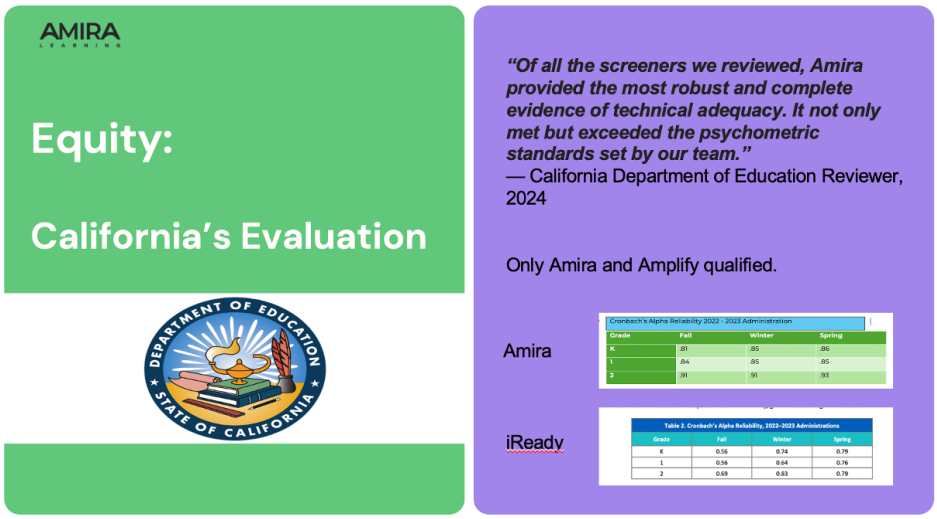
Universal Reading Screener	Overall Score (Max:155)	Constructs Measured (Max:65)	Technical Adequacy (Max:20)	Administration (Max:25)	Linguistic Diversity and Cultural Bias (Max:15)	Usability and Support (Max:30)
Amira ISIP	142.6	57.0	19.6	24.2	14.0	27.8
Star Universal Reading Suite	134.0	53.4	16.8	22.8	13.6	27.4
mCLASS with DIBELS 8th Edition	133.2	54.6	16.4	23.2	14.0	25.0
i-Ready Universal Screening Suite	120.2	50.4	14.4	20.8	11.0	23.6
MAP Reading Fluency	120.0	51.0	15.6	19.0	10.8	23.6
FastBridge Universal Reading Suite	118.0	45.6	16.0	20.8	12.0	23.6
Acadience Reading K-6	116.8	50	14.8	19.4	11	21.6
aimswebPlus	111.4	43	15.4	19.6	12.2	21.2

In an independent evaluation conducted by the Oklahoma State Department of Education under the Strong Readers Act (2025), Amira passed all 17 mandatory screening criteria and earned the highest score among all evaluated screeners on the full evaluation rubric — 266 points compared to 202 for DIBELS. The Oklahoma State Board of Education cited Amira's top marks for reliability and validity.

Oklahoma Evaluation Of Screeners This Month

Supplier Name	Name of Assessment	Mandatory Criteria	Non-Mandatory Criteria	Recommended for Approval
Amira Learning	Amira	Pass	266	Yes
Amplify Education	DIBELS 8th Edition	Pass	202	Yes
Curriculum Associates	i-Ready Diagnostic	Pass	201	Yes
Edmentum	Exact Path	Pass	201	Yes
EPS Learning	EPS Reading Assistant	Pass	210	Yes
HMH Learning – NWEA	MAP Growth Early Learning	Pass	210	Yes
NCS Pearson	aimswebPlus	Pass	155	Yes
Quest of OK – MIGHT	Literrific	Fail	---	No
Renaissance Learning	Star Early Learning	Pass	213	Yes
TouchMath Acquisitions	Classworks Reading Screener	Fail	---	No
UTJ Holdco–Teaching Strategies	GOLDFinch	Fail	---	No

In the 2024 California K-2 screener evaluation, Amira was the only screener unanimously approved by all expert reviewers across all grades in both English and Spanish.



State after state across the country has ranked Amira at the top or near the top for accuracy and fairness.

Summary of Section 1

Amira's equity is not an emergent property; it is an engineered one. The system was built from the ground up with a clear design philosophy (read errors, not accent), trained on intentionally diverse and representative data, annotated with explicit instructions to ignore accent-driven speech variation, monitored continuously for performance disparities, and independently validated for reliability by state psychometric teams nationwide.

2. Statistical Equity: Amira is equitable by the numbers

Design philosophy and training methodology are necessary but not sufficient. The proof of equity lies in the assessment's psychometric performance when disaggregated across student subgroups. This section presents the disaggregated reliability and item-fairness evidence from the Amira ISIP Assessment Technical Guide for the 2025–2026 school year.

Evaluating the fairness of any reading screener centers on two questions:

1. Does the assessment produce equally consistent and reliable scores across subgroups? (This is referred to as reliability.)
2. Do individual items perform similarly across student subgroups once you account for students' overall ability? (This is known as item fairness, measured through Differential Item Functioning (DIF).)

On both dimensions, Amira ISIP's performance meets or exceeds the highest standards in the K–3 screener market.

Reliability across student subgroups: the headline finding

Reliability coefficients measure the degree to which an assessment produces consistent, dependable scores. The widely accepted minimum standard for high-stakes screening decisions is a reliability coefficient of 0.80. Amira ISIP's reliability estimates meet or exceed 0.90 — well above the threshold — for virtually every grade-by-student subgroup combination evaluated.

The tables below summarize internal consistency reliability (as measured by Cronbach's alpha) and interrater reliability (Amira-human agreement) for the major demographic groups, drawn directly from the Amira ISIP Technical Guide.

Table 2.1. Reliability by gender (K–2)

Grade	Gender	N	Reliability	Agreement with Humans
K	Male	3,118	0.91	0.95
K	Female	3,067	0.91	0.95
1	Male	4,005	0.93	0.94
1	Female	3,851	0.93	0.94
2	Male	4,290	0.92	0.97
2	Female	4,110	0.93	0.97

Reading note. Internal consistency is essentially identical between boys and girls at every grade level, with no gap exceeding 0.01. Interrater reliability — the agreement between Amira's scoring and trained human scorers — is likewise identical by gender within grade.

Table 2.2. Reliability by English learner status (K–2)

Grade	Status	N	Reliability	Agreement with Humans
K	EL/MLL	1,026	0.91	0.97
K	Not EL/MLL	6,595	0.91	0.95
1	EL/MLL	1,120	0.96	0.91
1	Not EL/MLL	7,322	0.93	0.94
2	EL/MLL	1,154	0.93	0.98
2	Not EL/MLL	7,246	0.91	0.97

Reading note. Reliability for English learners is at parity with or exceeds reliability for non-English learners at every grade. In Kindergarten, EL/MLL interrater reliability (0.97) is higher than that of the non-EL/MLL comparison group (0.95). At Grade 1, EL/MLL internal consistency reaches 0.96 — the highest reliability cell in the entire EL comparison. There is no pattern of degraded performance for English learners.

Table 2.3. Reliability by race/ethnicity (K–2)

Grade	Race / Ethnicity	N	Reliability	Agreement with Humans
K	White	3,725	0.91	0.95
K	Hispanic	862	0.91	0.96
K	Asian	315	0.93	0.96
K	Black	1,653	0.90	0.94
K	Other	506	0.91	0.96
1	White	2,170	0.93	0.91
1	Hispanic	871	0.93	0.93
1	Asian	294	0.92	0.94
1	Black	1,662	0.93	0.94
1	Other	850	0.93	0.96
2	White	1,355	0.92	0.97
2	Hispanic	858	0.93	0.98
2	Asian	311	0.92	0.97
2	Black	1,765	0.92	0.98
2	Other	255	0.91	0.98

Reading note. Across 15 race-by-grade cells, every internal consistency reliability falls in the 0.90 -- 0.93 range, and every interrater reliability falls in the 0.91 -- 0.98 range. There is no race or ethnicity for which Amira's reliability drops below the high-stakes threshold of 0.80. In Grade 2, the interrater reliability for Hispanic, Black, and Other students (0.98, 0.98, 0.98) exceeds that for White students (0.97).

Table 2.4. Reliability by socioeconomic status (K–2)

Grade	SES	N	Reliability	Agreement with Humans
K	Low SES	500	0.90	0.98
K	High SES	964	0.90	0.97
1	Low SES	532	0.93	0.98
1	High SES	1,012	0.93	0.98
2	Low SES	575	0.92	0.98
2	High SES	949	0.90	0.97

Table 2.5. Reliability by home language (K–2)

Grade	Home Language	N	Reliability	Agreement with Humans
K	English	1,018	0.90	0.97
K	Spanish	446	0.90	0.97
1	English	1,252	0.91	0.98
1	Spanish	419	0.92	0.98
2	English	1,294	0.92	0.98
2	Spanish	246	0.91	0.98

Reading note. Socioeconomic and home language reliability are virtually indistinguishable across all comparisons. Low-SES students show higher or equivalent interrater reliability than high-SES students across all grades. Spanish home-language students post higher internal consistency than English home-language peers at Grade 1 (0.92 vs. 0.91). No student subgroup falls below the high-stakes threshold of 0.80.

Item-level fairness: Differential Item Functioning

Reliability alone is insufficient. Two student subgroups can show identical reliability while still exhibiting systematically biased individual items — items that are harder for one group

than the other, even after controlling for ability. The standard psychometric tool for detecting this is Differential Item Functioning, or DIF.

Amira ISIP uses IRT-based DIF analyses to flag items that perform differently across gender, Black/White, and Hispanic/White comparisons at the same level of underlying reading ability. The industry-accepted threshold for concern is an overall DIF flagging rate exceeding 5%, with a systematic pattern favoring one group. Amira ISIP's rates fall well below this threshold across every grade studied:

Table 2.6. Item-level DIF results across grades K–2

Grade	Comparison	Items with DIF	Total Items	DIF Rate
K	Gender (Male vs. Female)	2	124	1.61%
K	Black vs. White	6	124	4.84%
K	Hispanic vs. White	1	124	0.81%
1	Gender (Male vs. Female)	2	123	1.62%
1	Black vs. White	3	123	2.44%
1	Hispanic vs. White	5	123	4.00%
2	Gender (Male vs. Female)	1	100	1.00%
2	Black vs. White	5	100	5.00%
2	Hispanic vs. White	3	100	3.00%

Critical observations on the DIF results: First, all rates are at or below 5%, the conventional ceiling for unbiased operation. Second, the items flagged with DIF are spread across different subtests, with no systematic concentration in any single reading construct — the absence of a consistent pattern argues against systemic bias and supports statistical noise. Third, the small number of items showing DIF are split in both directions — some favor the focal group, some favor the reference group — again consistent with random variation rather than directional bias. Fourth, items flagged as DIF in any analysis are subsequently removed or recalibrated by Amira's psychometric team annually, ensuring that the active item bank remains DIF-clean over time.

AMIRA LEARNING

Summary of section 2

Across more than 60 reported reliability cells spanning gender, English learner status, exceptionality, race and ethnicity, SES, and home language, Amira ISIP reliability estimates almost universally meet or exceed 0.90, with no student subgroup falling below the 0.80 high-stakes threshold. DIF rates remain at or below 5% across all demographic comparisons at every grade level, and items showing DIF are removed or recalibrated annually. By the most demanding standards of psychometric equity, Amira ISIP is one of the most thoroughly validated reading screeners on the market.

3. Equity of Outcomes: Amira flags fewer minority students

The most direct test of whether Amira's scoring is unfair to English learners is to compare year-over-year proficiency outcomes in a state where the assessment instrument has changed, but the student population has not.

In the 2024 school year, New Mexico districts used the Istation Indicators of Progress (ISIP) assessment, an established multiple-choice and item-response screener — one that relied on keyboard-based tasks rather than on voice-based tasks (listening to children read aloud). In the 2025–2026 school year, New Mexico districts switched to Amira's ARM (Amira Reading Mastery) benchmark, which includes voice-scored oral reading as a core component of the total score.

If the central claim of voice-scoring critics is correct — that Amira's listening technology systematically disadvantages English learners — then English learner proficiency rates should have fallen compared to non-English learner rates when schools switched from the old keyboard-based test to Amira's voice-scored benchmark. And the effect should be clearest in Kindergarten and Grade 1, where Amira's voice tasks make up the largest share of the assessment. Those are the grades where a bias against EL students' speech would be most evident.

None of these predictions came true. In every case, the data shows the opposite.

What we'd expect to see if the claim were true

The predicted effects are specific and testable:

- The percentage of English learner students flagged as struggling readers (below the 20th percentile) should have increased, since the new tool introduces an additional source of error that would disproportionately affect English learner students.
- English learners should have been flagged at an even higher rate than before relative to non-English learners — the gap between the two groups should have widened.
- These effects should be observed across all grade levels, since Amira uses voice-scored tasks at every grade level.

Finding 1: Fewer English learners are being flagged as struggling, and the gap with non- English learner students has narrowed

The first test is straightforward: ***did more or fewer English-learner students be identified as struggling readers after switching to Amira***?

Under the old keyboard-based test, **54% of English learner students** in New Mexico were flagged as struggling readers (scoring below the 20th percentile at mid-year). Under Amira, that figure fell to **40%**. For non-English learner students, the rate also fell, from 29% to 24%.

Both groups improved — but English learners improved more. Under the old test, an English learner was 83% more likely to be flagged as struggling than a non-English learner. Under Amira, that figure dropped to 70% more likely to be flagged as struggling. The relative gap between EL and non-EL students narrowed by 7% — the opposite of what a voice-bias hypothesis predicts.

This result holds true regardless of how the data is analyzed. Whether using different assessment windows or the most recent data available for each student, the gap consistently narrows. Early end-of-year data show it continues to close further, to 34% above non-English learner rates, down from 83% under the old test.

Finding 2: English learners are reaching proficiency at much higher rates — and closing the gap with non-EL students

The second test does not look at who is being flagged as struggling, but at who is reaching proficiency. If Amira's voice scoring were suppressing English learner performance, proficiency rates should have fallen — particularly in Kindergarten and Grade 1, where voice-scored tasks make up the largest share of the assessment. Those are the grades where bias against EL students' accents would be most pronounced.

That is not what happened.

The table below shows the percentage of New Mexico students reaching proficiency on the mid-year assessment — separately for English learners and non-English learners — and how that comparison changed from last year (79,851 students on the previous keyboard-based tool) to this year (59,075 students on Amira's voice-scored benchmark).

Table 3.1. English learner (EL) vs. non-English learner (Non-EL) proficiency rates, mid-year, New Mexico

Grade	EL % Proficient Last Year	Non-EL % Proficient Last Year	How far behind?	EL % Proficient This Year	Non-EL % Proficient This Year	How far behind?	Did the gap narrow?
K	10.4%	31.4%	ELs at 33% of non-EL rate	16.7%	40.3%	ELs at 41% of non-EL rate	✓ Yes

Commented [1]: we need formatting and labels that make this understandable. @dianne.henderson@amiralearning.com can someone run this thru Claude and just make some judgments. _Assigned to dianne.henderson@amiralearning.com_

Commented [2R1]: revised @chris.blevins@amiralearning.com @mark.angel@amiralearning.com @ran.liu@amiralearning.com did I miss anything important?

Grade	EL % Proficient Last Year	Non-EL % Proficient Last Year	How far behind?	EL % Proficient This Year	Non-EL % Proficient This Year	How far behind?	Did the gap narrow?
1	13.0%	31.4%	ELs at 41%	21.0%	35.6%	ELs at 59%	✓ Yes
2	20.5%	35.0%	ELs at 59%	27.7%	41.5%	ELs at 67%	✓ Yes
3	19.0%	37.6%	ELs at 51%	27.5%	45.9%	ELs at 60%	✓ Yes
4	10.1%	28.2%	ELs at 36%	25.6%	42.1%	ELs at 61%	✓ Yes
5	8.6%	30.9%	ELs at 28%	22.5%	38.7%	ELs at 58%	✓ Yes
Overall	**14.7%**	**32.7%**	**ELs at 45%**	**24.6%**	**40.1%**	**ELs at 61%**	✓ Yes

How to read the "How far behind?" column. If English learners and non-English learners reached proficiency at exactly the same rate, this number would be 100%. Last year, statewide, English learners were reaching proficiency at only 45% of the rate of their non-English learner peers — less than half. This year, under Amira, that figure is 61%. The gap is still real, but it has narrowed substantially at every single grade.

What does "36% improvement" mean? The gap moved from 45% to 61% — an improvement of 16 points. Relative to where we started (45%), that 16-point gain represents a 36% reduction in the size of the disadvantage. English learners have closed more than a third of the ground they were behind.

The youngest grades tell the clearest story. If Amira's voice scoring were harder on English learners, Kindergarten and Grade 1 — the most voice-intensive grades — should show the worst results. Instead, Grade 1 shows the single largest improvement of any grade: the proficiency gap narrowed by 18 points. Kindergarten improved by 8 points. The grades that critics predicted would decline actually improved. This is inconsistent with a tool that cannot fairly hear EL students.

What stands out in the New Mexico data

Four findings are directly relevant to the question of voice-scoring fairness.

Every grade narrowed — including the voice-heaviest grades

If voice scoring systematically penalized English learners, the bias signature would be strongest in the early elementary grades, where voice-heavy tasks dominate the construct. Kindergarten and Grade 1 should show the largest year-over-year regressions in ELL parity. They show some of the greatest improvements. Grade 1 ELL parity improved by 0.18 ratio points, the second-largest improvement of any grade. Kindergarten improved by 0.08 ratio points.

Absolute ELL proficiency rates were much higher with Amira

This is not a story about non-ELL proficiency falling and making the ratio look better by comparison. ELL proficiency rates rose substantially in absolute terms: Grade 4 ELL proficiency rose from 10.1% to 25.6% (a 2.5× multiplier), Grade 5 ELL proficiency rose from 8.6% to 22.5% (a 2.6× multiplier), and Grade 1 ELL proficiency rose from 13.0% to 21.0% (a 1.6× multiplier). These are large absolute gains, not consistent with the hypothesis that the new voice-scored instrument is suppressing ELL performance.

ELLs are improving faster than non-ELLs in relative terms

ELL proficiency rose by approximately 10 percentage points statewide (from 14.7% to 24.6%). Non-ELL proficiency rose by approximately 7 percentage points (from 32.7% to 40.1%). The parity ratio moved from 0.45 to 0.61 — a 36% reduction in the relative proficiency gap. Whatever else is happening in New Mexico, English learners are closing ground on their non-English-learner peers under the new instrument, not losing ground.

The narrowing is robust to the MOY-window definition

Methodological skeptics rightly ask whether the result depends on how the analyst slices the data. It does not. Every defensible analytical cut of the 2025–2026 data shows substantial closure of the proficiency gap relative to the 2024 ISIP ratio of 0.45: P5-only ARM benchmark window (ratio = 0.68), P4-only ARM benchmark window (ratio = 0.60), the latest of P4 or P5 headline analysis (ratio = 0.61), and P8 EOY with partial coverage (ratio = 0.75). The headline ratio of 0.61 is actually one of the more conservative cuts. The improvement holds across every window definition tested.

What this result shows

This year-over-year analysis directly refutes the specific claim that voice tasks and Amira scoring have preferentially harmed English learners in New Mexico. The predicted signature of voice-induced bias is unambiguous: ELL proficiency falling relative to non-ELL proficiency. The opposite is observed at every grade level under every defensible cut of the data.

Two caveats are worth stating plainly. First, this analysis does not by itself prove that ARM voice scoring is bias-free in general — comprehensive bias evaluation requires per-task fairness audits, accent-disaggregated error analyses, and continued monitoring of demographic outcomes throughout the school year (work that is ongoing). Second, ARM grade calibration is still being refined, which may shift absolute proficiency numbers (though not relative ratios) in future analyses. The fundamental finding — that ELL parity improved substantially under the new voice-scored instrument — is robust to these refinements.

Summary of section 3

The transition from a legacy multiple-choice screener to Amira's voice-scored ARM benchmark in New Mexico was accompanied by a 36% reduction in the relative proficiency gap between English learners and non-English learners. Every grade narrowed. Absolute ELL proficiency rates doubled or tripled. The narrowing is robust to every defensible analytical cut. This pattern is the opposite of what the voice-bias hypothesis predicts.

4. Explaining Perceptions: When scores look wrong, the cause is invariably audio quality

Amira is deployed at scale across millions of student sessions per year. Across that volume, some sessions inevitably produce scores that look implausible to a teacher familiar with the student. These cases warrant investigation. The evidence from tens of thousands of such investigations over the years is unequivocal: the overwhelming majority of these cases are explained by audio-capture conditions or protocol violations rather than by limitations of the underlying Amira model.

This section presents the most recent district-level scoring validation evidence, summarizes the implementation conditions required for accurate evaluation, and explains why audio capture issues account for the remaining disagreements after those conditions are met.

The Texas ISD spring 2026 scoring validation study

In spring 2026, Amira's psychometrics team partnered with Texas ISD on a rigorous human-versus-Amira scoring validation study. The study was designed to meet the highest standards of academic research and to deliver results that district leaders could present to their board with confidence.

Study design

- **Sample.** Seven Texas ISD schools participated, with demographics from the March progress monitoring window aligning with district-wide middle-of-year demographics. The sample included a balanced representation of emergent bilingual (EB) and non-EB classrooms.
- **Protocol.** Human scorers evaluated student oral reading fluency (ORF) sessions using a standardized, rubric-based scoring protocol. Scores from Amira's production model were then compared directly with human scores at the word and activity levels.
- **Expert review.** Results were independently reviewed and validated by Amira's Vice President and Chief AI Scientist.
- **Transparency.** Findings were shared directly with district leaders, providing full transparency into Amira's scoring performance across all student groups.

Headline results

Metric	Value	Interpretation
ORFs compared	1,827	Oral reading fluency activities across seven schools, K–5
Total words scored	301,858	Individual words assessed and compared
Overall accuracy	96.9%	Human-to-Amira word-level agreement
Median per-activity accuracy	98.17%	Median activity-level accuracy
Average Cohen's Kappa	0.6977	A statistical measure of agreement. Anything above 0.6 is considered substantial agreement by researchers.

Published research in this domain consistently shows 96–97% agreement between trained human raters and well-calibrated Amira scoring systems. Texas's overall accuracy of 96.9% is squarely within the published inter-rater reliability standards — in fact, slightly above the midpoint of the established research range.

Results by first language

The most direct test of equity in a voice-scored assessment is whether scoring accuracy holds for students whose native language is not English. The Texas study disaggregated results by a student's first language and produced the following:

Native Language	N	Total Words	Overall Accuracy	Median Accuracy	Cohen's Kappa
English (EN)	1,315	222,236	97.2%	98.25%	0.7058
Non-EN (ES, OT, ZH, RU, etc.)	469	73,185	95.8%	97.66%	0.6759
No Roster Match (NaN)	43	6,437	97.1%	98.32%	0.6861

Reading the result. Non-English native language students achieved 95.8% scoring accuracy — right around Amira's researched 95–97% standard — with a Cohen's Kappa of

0.6759, well above the 0.6 threshold for substantial agreement. The 1.4-percentage-point difference between English- and non-English-native language groups is within the noise floor of any comparable inter-rater reliability study. There is no evidence in this dataset of a meaningful penalty associated with non-English native-language status.

Why scoring disagreements happen: the audio-quality root cause

When Amira's scoring disagrees with a human rater on a particular session, the source of the disagreement is almost always traceable to a small number of root causes — and none of them is the Amira model's underlying accuracy. Amira RED is intended for real-life classroom use and depends on a protocol that mirrors real-world implementations. When that protocol is followed, accuracy reaches the levels reported above. When it is not, accuracy degrades — not because the model is biased, but because the input it receives is degraded.

Implementation conditions for valid scoring evaluation

The Amira RED equity white paper specifies five implementation conditions that must be met for any meaningful evaluation of scoring accuracy. These conditions are not arbitrary requirements; they are direct consequences of how RED is designed.

- **One-to-one student-to-login mapping.** Amira RED adapts dynamically to each student's speech context. Using a single login across many different students and voices degrades scoring accuracy because the model's adaptation is being misdirected. Every student must have their own login.
- **Accurate rostering.** Amira RED models use rostering inputs (including grade level) as features in the scoring process. Inaccurate roster data — a third grader logged in as a first grader, for example — will produce inaccurate scores, not because the model is wrong but because it is operating on wrong inputs.
- **Adherence to environmental rules.** Scoring accuracy is materially affected by the audio environment: physical spacing among test-takers, ambient classroom noise, and the use of headphones where appropriate. Amira flags low-quality audio sessions (the "yellow flag" system) and recommends specific sessions for teacher review and potential re-testing. Failing to act on these recommendations will produce a distorted picture of overall accuracy.
- **Proper understanding of Amira RED's design.** Evaluators sometimes treat Amira as if it were a standard ASR system, a mispronunciation detector, or an alternative dialect-proof translation engine. It is none of these. Amira's job is to evaluate

whether a student is reading the phonemes in an English word in a way consistent with a teacher-normed expectation. It does not judge pronunciation purity, does not require an American accent, and does not completely discount accent or dialect when those features are indistinguishable from a reading error.

- **No conflation of assessment and practice modes.** Amira deploys two distinct RED models: one for assessment, optimized for measurement of reading mastery; and one for practice, optimized for real-time micro-intervention. They have different latency and feature constraints, different output characteristics, and different intended uses. Evaluating an assessment session against a practice model (or vice versa) will produce misleading conclusions.

Two illustrative cases from the Texas study

Two cases from the Texas study illustrate the audio-issue pattern directly. In the first, a student's session shows clear off-task behavior — the student is not engaged with the reading passage and is producing speech unrelated to the target text. In the second, an English learner session shows poor audio quality — the headphone microphone is positioned incorrectly, ambient classroom noise dominates the signal, and the model is operating on degraded input. In both cases, the resulting score is not representative of the student's underlying reading ability. In neither case is the Amira scoring model itself the source of the problem.

Sessions like these are precisely what Amira's yellow-flag protocol is designed to identify. When the protocol is followed and flagged sessions are reviewed by teachers, the residual scoring accuracy is the 96–97% figure reported in the Texas study and in published research. Districts conducting their own scoring validation studies should confirm one-to-one student-to-login assignment, verify accurate rostering, use headphones in noisy environments, train evaluators on the distinction between reading errors and accent variation, and never compare assessment-mode and practice-mode sessions across the two RED models.

Summary of section 4

When proper implementation conditions are met, Amira's scoring accuracy reaches 96.9% overall and 95.8% for non-English native language students — squarely within published research standards for inter-rater reliability. The relatively rare cases where Amira's scoring disagrees materially with human scoring are almost always traceable to audio capture issues, rostering errors, or protocol violations rather than to any underlying limitation of the Amira scoring model.

5. Scoring Equity: Testing shows no disparity for students of color and English learners

The most direct equity claim a voice-scored reading assessment can make is that its scoring accuracy does not vary meaningfully by student demographic group. This section presents the evidence behind that claim, drawn from Amira's held-out test set and the broader Reading Error Detection equity white paper.

Test set results

Amira's end-to-end Reading Error Detection model is trained on expert-labeled data from thousands of students.

Table 5.1. Reserved test set: aggregate scoring accuracy

Group	Word-Level Accuracy	% of Test Set
All students (overall)	95.1%	100%

Table 5.2. Reserved test set: scoring accuracy by student group

Student Group	Word-Level Accuracy	Delta vs. Overall	% of Test Set
African American	95.6%	+0.5%	25%
English Language Learner (MLL/ELL)	95.1%	0.0%	28%
Other U.S. Students	95.0%	-0.1%	33%
International (non-U.S.)	94.2%	-0.9%	14%

The headline finding....

Amira's word-level scoring accuracy for African American students is higher than its accuracy for the overall population (95.6% vs. 95.1%). Accuracy among English learners is statistically identical to that of the overall population. Accuracy for Other U.S. students is within 0.1% of the overall accuracy. The only student subgroup showing any meaningful

deviation — international students, at 94.2% — is the one with the largest accent and dialect variance, and the deviation is less than 1 percentage point. There is no pattern of degraded accuracy for students of color or English learners.

WCPM correlation: end-to-end equity at the score level

Word-level accuracy is one way to measure scoring fidelity. The more practically meaningful question is whether Amira and human scorers produce equivalent aggregated assessment scores, since the aggregated score is what drives classroom instructional and intervention decisions. The relevant overall score for oral reading fluency is Words Correct Per Minute (WCPM), which measures how many words a student reads correctly per minute.

The table below reports the correlation between Amira's WCPM scores and human-scorer WCPM scores, both overall and broken down by student group. A correlation of 1.00 would mean perfect agreement; anything above 0.90 is considered exceptionally strong.

Table 5.3. Human-Amira WCPM correlation, held-out test set

Group	Pearson r	Interpretation
All students	0.995	Near-perfect correlation with human scoring
African American	0.990	Near-perfect correlation
English Language Learners (ELLs)	0.997	Higher than overall correlation
Other U.S. Students	0.997	Higher than overall correlation
International (non-U.S.)	0.990	Near-perfect correlation

Reading the correlation table. Every student subgroup correlation is at or above 0.99. The English learner group and the Other U.S. student group both post correlations of 0.997 — marginally higher than the overall correlation of 0.995. The African American and international groups both post correlations of 0.990, within five thousandths of the overall figure. By the most demanding standard of end-to-end equity — do Amira and human scorers produce equivalent assessment outcomes for every student subgroup? *The answer is yes.*

Why Amira's student subgroup accuracy can match (or exceed) overall accuracy

A naive expectation might be that scoring accuracy must degrade for students with accents, dialects, or non-native English speech patterns. That expectation is correct for standard ASR systems, which try to transcribe open-ended speech and are penalized whenever the speaker's acoustic patterns deviate from the dominant training distribution. It is not correct for Amira.

Three design choices explain why Amira's student subgroup accuracy does not degrade in the way ASR systems do:

- **Grounded in target text, not in open-ended speech.** Because Amira already knows what the student is supposed to read, it does not perform open-ended speech-to-text conversion. It only determines whether the student's production of each target phoneme is consistent with reading mastery. This dramatically reduces the surface area where accent and dialect could cause problems.
- **Training data weighted toward accent- and dialect-bearing populations.** Half of the speech-layer training corpus comes from schools with 75%+ students-of-color enrollment. The expert-labeled training set is 25% African American, 28% English learner, and 14% international. This overfitting ensures the model has seen ample examples of the speech patterns it is most likely to encounter in production.
- **Annotators instructed to ignore consistent accent or dialect patterns.** Reading miscues are labeled only when the speech deviation indicates a decoding failure, not when it reflects a stable accent or dialect feature. The model is trained, by design, to behave the way teachers expect: distinguishing reading errors from speech variation.

Continuous monitoring: equity as an ongoing process

The goal of minimizing bias in an Amira system is never "achieved" — it is maintained through continuous monitoring. Amira's ongoing equity work includes per-iteration student subgroup performance audits (every new RED model iteration is evaluated by disaggregating performance metrics across student subgroups, plus age and overall reading level, with degradation triggering remediation), annual DIF analyses using RED-model-scored item responses as inputs (so any bias propagating from the speech scoring layer would surface in DIF results), independent state agency review (which has led to Amira's approval as a state-recognized screener in nearly every state where such review processes exist), and annual item-bank refresh.

AMIRA
LEARNING

Summary of section 5

Amira's scoring accuracy for African American students (95.6%) exceeds its accuracy for the overall student population (95.1%). Scoring accuracy among English learners is statistically identical to that of the overall population. WCPM correlations with human scorers are at or above 0.99 for every student subgroup. By every measure of equity that the Research Field has developed — word-level accuracy, item-level fairness, end-to-end score correlation, and formal interrater reliability — Amira's scoring works equally well, or better, for students of color and English learners than it does for the overall population.

6. Comparative Equity: Other screeners are less equitable

The evidence presented in this paper shows that Amira assesses English learners, students of color, and students with disabilities fairly and accurately. A reasonable follow-up question is whether other reading screeners do better. Based on what those screeners have submitted to state education agencies across multiple states, the independent evaluations of the National Center on Intensive Intervention, and the Buros Center for Testing, the answer is no.

Over the past several years, states have increasingly required reading screeners to submit their measurement data broken out by student group as a condition of approval. The most comprehensive evidence comes from three sources: California's 2024 K-2 screener competitive evaluation; Colorado's 2022 READ Act competitive submissions, independently scored by researchers at the University of Massachusetts Amherst; the Buros Center for Testing's January 2026 psychometric review commissioned by the Nebraska Department of Education; the WestEd 2024 Massachusetts Early Literacy Screening Analysis; and NCII, the national standards body that evaluates reading screeners for use in federal improvement programs. Taken together, these reviews represent the most systematic cross-vendor comparison of equity evidence ever conducted in the K-3 literacy screener market.

The picture that emerges from all five sources is consistent: most screeners have not collected equity data at a meaningful scale, have data showing problems they chose not to publish, or have submitted nothing at all for key student groups.

What the competitive data shows

The table below summarizes what the three major screeners reported by the student group. A score of 0.80 is the widely accepted minimum for a tool used to make high-stakes screening decisions about children. A score of 0.90 is the standard required for individual educational placement decisions.

Table 6.1. Subgroup Reliability, Interrater Reliability (IRR), Test-Retest (TR), and Alternate Form (AF) Data

*(Metrics Key: **R** = Internal Consistency/Marginal Reliability; **IRR** = Interrater Reliability; **TR** = Test-Retest Reliability; **AF** = Alternate/Parallel Form Reliability; **DIF** = DIF Analysis Conducted)*

Grade	Student Group	Amira	DIBELS (CBM)	STAR (CAT)
K	**All students**	R: 0.86–0.91 TR: .84, DIF ✓	R: 0.73–0.84, TR: .72–.92	R: 0.78–0.86, TR: .64
	English Learners	R: 0.91, IRR: .97, DIF ✓	Not provided Not provided DIF ✓	Not provided Not provided
	Students with Disabilities	R: 0.90, IRR: .96, DIF ✓		—
	Hispanic/Latino	R: 0.91, IRR: .96, DIF ✓	R: 0.70	—
	Black/African American	R: 0.90, IRR: .94, DIF ✓	Not provided Not provided DIF ✓	—
	Low income	R: 0.90, IRR: .98, DIF ✓	DIF ✓	—
	Home language: Spanish	R: 0.90, IRR: .97	Not provided	Not provided
1	**All students**	R: 0.93–0.96, TR: .86, DIF ✓	R: 0.89–0.91, TR: .74–.92	R: 0.78–0.86, TR: .61
	English Learners	R: 0.96, IRR: .91, DIF ✓	Not provided Not provided DIF ✓	Not provided
	Students with Disabilities	R: 0.92, IRR: .96, DIF ✓	Not provided Not provided DIF ✓	Not provided

Grade	Student Group	Amira	DIBELS (CBM)	STAR (CAT)
	Hispanic/Latino	R: 0.93, IRR: .93, DIF ✓	R: 0.91,	Not provided
	Black/African American	R: 0.93, IRR: .94, DIF ✓	Not published	Not provided
	Low income	R: 0.93, IRR: .98, DIF ✓	Not provided Not provided DIF ✓	Not provided
	Home language: Spanish	R: 0.92, IRR: .98	Not published	Not provided
2	**All students**	R: 0.91–0.93, TR: .87, DIF ✓	R: 0.87–0.89, TR: .57–.88	R: 0.78–0.86, TR: .79
	English Learners	R: 0.93, IRR: .98, DIF ✓	Not provided Not provided DIF ✓	Not provided
	Students with Disabilities	R: 0.92, IRR: .98, DIF ✓	Not provided Not provided DIF ✓	Not provided
	Hispanic/Latino	R: 0.93, IRR: .98 DIF ✓	R: 0.89,	Not provided
	Black/African American	R: 0.92, IRR: .98, DIF ✓	Not provided Not provided DIF ✓	Not provided
	Low income	R: 0.92, IRR: .98, DIF ✓	Not provided Not provided DIF ✓	Not provided

Grade	Student Group	Amira	DIBELS (CBM)	STAR (CAT)
	Home language: Spanish	R: 0.91, IRR: .98	Not provided	Not provided

Sources: California K-2 screener competitive submissions (2024); Colorado READ Act FOIA submissions (2022), University of Massachusetts Amherst review; Buros Center for Testing Nebraska NRIA Review (January 2026); NCII Academic Screening Tools; vendors' published technical manuals. "—" = no subgroup data submitted. Minimum screening standard: 0.80.

What five independent reviews found

Finding 1: Amira achieved NCII's full compliance ratings across all technical standards. NCII — the national benchmark used by federal improvement programs — awarded full ratings to only two of the major K-3 tools. DIBELS currently holds "Partially Met" ratings on multiple NCII technical categories. STAR's overall reliability of 0.78–0.86 falls below the 0.90 threshold required for individual placement decisions. NWEA MAP carries "more evidence needed" designations for subgroup data and K-1 predictive validity.

Finding 2: The Buros Center's independent Nebraska review flagged fairness gaps for nearly every competitor. In January 2026, the Nebraska Department of Education commissioned the Buros Center for Testing to independently review 11 screeners from seven vendors. The Buros team — a panel of national psychometric, screening, and reading experts — found that missing or incomplete fairness evidence was a specific weakness cited for DIBELS, FastBridge, i-Ready, and others. Acadience's review was discontinued entirely due to outdated norms. MAP Reading Fluency was rated "more evidence needed" on both appropriateness and technical adequacy. DIBELS' weaknesses were specifically noted to include classification accuracy, classification consistency, and fairness.

Finding 3: DIBELS' Kindergarten reliability falls below standard for Hispanic students, and it published nothing for English learners. DIBELS' own California submission shows a reliability score of 0.70 for Hispanic Kindergartners — below the 0.80 minimum. At Grades 1 and 2, the scores improve to 0.91 and 0.89, still below Amira's 0.93 at those grades. For English learner students and students with disabilities at every grade, DIBELS noted only that DIF analyses were conducted, but published no reliability data. The Colorado review independently confirmed this pattern: University of Massachusetts researchers scored Amplify a 1 out of 3 on EL accommodation evidence — and gave the same score to five of the six other vendors.

Finding 4: STAR and NWEA published no subgroup data, and WestEd's independent state analysis flags STAR as a potential over-identifier for ELLs. Across California, Colorado, Nebraska, and NCII reviews, neither STAR nor NWEA submitted reliability data broken out by EL status, disability status, race, income, or home language. The blank cells in the table above represent every grade and every subgroup. Beyond the absence of data, WestEd's 2024 Massachusetts Early Literacy Screening Analysis — an independent review of statewide screening data — found that Renaissance/STAR flags 25–33% of students overall and potentially up to 40% of English learners as at-risk, which WestEd described as a potential "cultural bias in identification" risk. By comparison, Amira's flag rate of 15–17% is calibrated to the IDA's estimated prevalence of neurological dyslexia, minimizing the false-positive burden that disproportionately falls on EL students.

Why sample size matters

The reliability numbers only tell part of the story. The equally important question is: how many students of each type were in the study? Amplify's California submission showed approximately 5% of students as English learners per grade — with English Learner status completely missing or unreported for nearly half (48%) of the Kindergarten sample — and approximately 16 to 18% Hispanic students. New Mexico's student population is approximately 16% English learners and nearly half Hispanic. A screener validated on a sample that was overwhelmingly majority-language students offers no documented assurance of equitable performance for the students who make up the majority of New Mexico classrooms.

Amira benchmarked its entire validation study to match the actual makeup of the student populations it serves — including 24.5% English learners, 25.1% students with disabilities, and 50.1% Hispanic students — and ensured a minimum of 2,000 students per subgroup, so that every reported result is statistically meaningful for each group. For Spanish-speaking students specifically, Amira also developed EL-specific national norms using data from more than 52,000 Spanish-English bilingual students, allowing teachers to compare EL students against relevant peers rather than the general population.

How Amira was built for equity — not retrofitted for it

Amira's equity performance is not the result of patches applied to a standard assessment. It is the product of deliberate architectural choices made before a single student was tested.

Unlike tools that adapt English assessments for Spanish-speaking students through translation, Amira's Spanish assessment — Amira Evaluar — was developed from the ground up with native Spanish-speaking educators and is aligned to the specific phonological and

orthographic structure of Spanish. Spanish is a vowel-centered language with transparent orthography, and Amira's Spanish tasks reflect that — focusing on syllable segmentation, blending, and manipulation rather than English phonics conventions. Spanish-speaking students are evaluated on what actually predicts reading success in Spanish, not on a translated proxy.

Having both English and Spanish assessments allows educators to build a cross-language profile for each student. When a student struggles in English, the Spanish results help answer a question that traditional screeners cannot: Is this a reading difficulty, or a language acquisition gap? Getting that distinction right determines whether a child is referred for reading intervention, English language support, or both. Amira is designed to reliably make that distinction. For EL students taking the English assessment, Amira also provides Spanish-language proctoring so students understand what they are being asked to do, regardless of English proficiency.

The contrast is not a reporting difference. Other screeners were built for a general population, and their equity evidence reflects that. Amira was designed for the full range of students in American classrooms — including New Mexico's.

Conclusion

Equity in Amira-powered assessment is a design discipline, not a marketing claim. It is established through the convergence of architectural decisions, training data composition, annotation protocols, item-level fairness analyses, year-over-year outcome monitoring, and independent district- and state-level validation studies. By each of these standards, Amira's evidence base is among the most thorough in the K–3 literacy assessment market.

The five lines of evidence assembled in this paper converge on a consistent picture. Amira is designed to detect reading skill gaps, not accent or dialect variation — grounded in target text, not in open-ended speech transcription. Amira's psychometric reliability and item-level fairness hold across all measured student subgroups, with reliability coefficients at or above 0.90 in nearly every cell and DIF rates uniformly at or below 5%. In the most direct year-over-year test — New Mexico's transition from a multiple-choice screener to Amira's voice-scored ARM benchmark — English learner proficiency parity with non-English learners improved by 36%, with every grade narrowing and absolute ELL proficiency rates doubling or tripling.

Where third-party scoring concerns do emerge, they are traceable to audio capture issues, rostering errors, or protocol violations rather than to underlying Amira scoring limitations. When implementation conditions are met — as in the Texas ISD spring 2026 validation study — Amira's scoring agrees with trained human scorers at 96.9% overall and 95.8% for non-English native language students, squarely within published research standards. Across every direct measure of student subgroup equity, Amira performs as well or better for students of color and English learners than for the overall population: African American scoring accuracy (95.6%) exceeds overall accuracy, English learner scoring accuracy (95.1%) equals overall accuracy, and WCPM correlations with human scorers are at or above 0.99 for every student subgroup studied.

Amira is committed to maintaining this evidence base through ongoing per-iteration student subgroup audits, annual DIF analyses, state agency reviews, and continuous item-bank refresh. The work of equity in the Amira assessment is never complete. The work to date, however, supports a clear conclusion: when Amira-powered reading assessment is designed for equity from the ground up, trained on representative data, annotated with the right discipline, and continuously monitored across the populations it serves, it can deliver more equitable outcomes than the assessments it replaces.

AMIRA
LEARNING

For more information

Amira Learning publishes ongoing evidence of equity and validity through the Amira ISIP Technical Guide, district-level validation study reports, and state agency review submissions. For specific evidence relating to a state's adoption process or a district's validation question, contact the Amira Learning team at info@amiralearning.com.

Cited Sources

- Amira ISIP Assessment Technical Guide, School Year 2025–2026, sections 4.6 (Differential Item Functioning), 9.1.4 (Inter-rater Reliability), and 9.1.5 (Reliability of student subgroups).
- Amira Reading Error Detection: Evidence of Equity and Inclusion (RED Assessment Model white paper, 2026).
- Amira Scoring: Built for Accuracy and Fairness (Amira Learning product overview, 2026).
- Effects of Culturally Congruent Educational Technologies on Student Achievement (cited in Amira RED equity white paper).
- Jonson, J.L., Moses, T., Kettler, R., Bazis, P., & Fisher, E. (January 2026, amended January 26, 2026). *Psychometric Review of Reading Screeners for the Nebraska Reading Improvement Act*. Buros Center for Testing, University of Nebraska-Lincoln. Commissioned by the Nebraska Department of Education. https://www.education.ne.gov/nebraskareads/approved-assessments/
- Koenecke, A., et al. (2020). Racial disparities in automated speech recognition. Proceedings of the National Academy of Sciences.
- Lemke, M., Murphy, D., Chow, A.S.P., Spencer, H., & Zhang, A. (2023). *A First Look at Early Literacy Performance in Massachusetts: Results of Initial Analysis Based on State Grantee Literacy Screening Assessments*. WestEd. https://eric.ed.gov/?id=ED638580
- National Center on Intensive Intervention. (continuously updated). *Academic Screening Tools Chart*. American Institutes for Research, funded by the U.S. Department of Education Office of Special Education Programs. https://charts.intensiveintervention.org
- New Mexico Year-over-Year Proficiency Analysis: Did Voice Tasks and Amira Scoring Hurt ELL Students? (R. Liu, internal Amira Learning analysis, May 2026).

- Oklahoma State Department of Education, Report on Literacy Screener Evaluations (2025). https://oklahoma.gov/content/dam/ok/en/osde/strong-readers-files/Report%20on%20Literacy%20Screener%20Evaluations%202025.pdf
- Texas ISD Spring 2026 Scoring Validation Study (Amira Learning Psychometrics Team).