Academic Screening Tools Chart
Rating Rubrics: Spanish Tools

*Please note that the following rubrics are applied separately for **each grade level** targeted by the tool.*

Technical Standard 1. Classification Accuracy

Note: Classification Accuracy will be rated separately for each time of year for the administration (e.g., Fall, Winter, Spring).

Rating	Definition
Full Bubble	All of Q1 – Q4 rated as YES ***and*** The lower bound of the confidence interval around the Area Under the Curve (AUC) estimate ≥ 0.80 ***and*** Sensitivity ≥ 0.80 and Specificity ≥ 0.80
Half Bubble	All of Q1 – Q4 rated as YES ***and*** (a) The lower bound of the confidence interval around the AUC estimate ≥ 0.70 but < 0.80 ***or*** (b) Sensitivity ≥ 0.70 and Specificity ≥ 0.70
Empty Bubble	Does not meet full or half bubble
Dash	Classification accuracy data were not provided

Q1. Was an appropriate external measure of academic performance administered in Spanish used as an outcome?

Q2. Was the language of instruction of the sample used for classification analyses consistent with the language of the tool?

Q3. Was risk adequately defined within an RTI approach to screening (i.e., 10^{th}-20^{th} percentile), and consistent with the language of instruction?

Q4. Were the classification analyses and cut-points adequately performed?

Area Under the Curve (AUC) Statistic: an overall indication of the diagnostic accuracy of a Receiver Operating Characteristic (ROC) curve. ROC curves are a generalization of the set of potential combinations of sensitivity and specificity possible for predictors. AUC values closer to 1 indicate the screening measure reliably distinguishes between students with satisfactory and unsatisfactory reading performance, whereas values at 0.50 indicate the predictor is no better than chance.

Technical Standard 2: Reliability

Rating	Definition
Full Bubble	(a) A model-based approach to reliability was reported ***or*** (b) At least two types of reliability were reported that are appropriate for the purpose of the tool (e.g., inter-rater reliability is provided for tools that require human judgment), and evidence is drawn from at least two samples that are representative of heritage Spanish-speaking students across all performance levels ***and*** For each type of reliability reported, the median lower bound of the confidence interval around the estimate met or exceeded 0.70.
Half Bubble	(a) A model-based approach to reliability was reported ***or*** (b) At least two types of reliability were reported that are appropriate for the purpose of the tool (e.g., inter-rater reliability is provided for tools that require human judgment), and evidence is drawn from at least one sample that is representative of heritage Spanish-speaking students across all performance levels ***and*** For each type of reliability reported, the median lower bound of the confidence interval around the estimate met or exceeded 0.60.
Empty Bubble	Does not meet full or half bubble.
Dash	Reliability data were not provided.

Technical Standard 3: Validity

Rating	Definition
Full Bubble	At least two types of appropriately justified* validity analyses are reported, ***and*** the analyses are drawn from at least one sample representative of heritage Spanish-speaking students across all performance levels, ***and*** the median lower bound of the confidence interval around the estimate met or exceeded 0.60 (or was within an acceptable range given the expected relationship with the criterion measure).
Half Bubble	Analyses, measures, and sample were appropriate for at least two types of appropriately justified validity analyses*, but evidence was mixed, with the median lower bound of the confidence interval for one or more, but not all, types of validity either not meeting or exceeding 0.60 or not within an acceptable range given the expected relationship with the criterion measure.
Empty Bubble	Does not meet full or half bubble.
Dash	Validity data were not provided.

**Appropriately justified analyses must include at least one criterion measure that is external to the screening system and theoretically linked to the underlying construct measured by the tool.*

Technical Standard 4: Sample Representativeness

Description	Definition
National with Cross-Validation	At least one classification accuracy analysis was conducted using a national sample* ***and*** at least one cross-validation study was conducted.
National without Cross-Validation	At least one classification accuracy analysis was conducted using a national sample* ***without*** a cross-validation study.
Regional with Cross-Validation	At least one classification accuracy analysis was conducted using one or more state or regional samples ***and*** at least one cross-validation study was conducted.
Regional without Cross-Validation	At least one classification accuracy analysis was conducted using one or more state or regional samples ***without*** a cross-validation study.
Local with Cross-Validation	At least one classification accuracy analysis was conducted using one or more local district samples ***and*** at least one cross-validation study was conducted.
Local without Cross-Validation	At least one classification accuracy analysis was conducted using one or more local district samples ***without*** a cross-validation study.

**A national sample consists of at least 150 students across at least three of nine geographical divisions defined by U.S. Census Bureau.*

Technical Standard 5: Bias Analysis

Bias Analysis refers to an analysis that examines the degree to which a tool is or is not biased against subgroups (e.g., race/ethnicity, gender, socioeconomic status, students with disabilities, English language learners).

Rating	Definition
Yes	Analyses testing differential classification accuracy across demographic and language proficiency groups were conducted. Other bias analyses that can be reported as an addition to differential classification accuracy include: 1. Multiple-group confirmatory factor models for categorical item responses 2. Explanatory group models such as multiple-indicators, multiple-causes (MIMIC) or explanatory IRT with group predictors 3. Differential Item Functioning from Item Response Theory (DIF in IRT)
No	Does not meet “yes”

Technical Standard 6: Item Development

Rating	Definition
Full Bubble	Item development – including writing processes, item writer and reviewer training and knowledge, and item review processes for judging item appropriateness – indicates that items adequately account for the measured construct, cultural prototypicality, Spanish language development, syntax, dialect representation and variation, and the age of acquisition of skills or words.
Half Bubble	Item development indicates that items appropriately account for the measured construct, but mixed evidence is provided that the items adequately account for cultural prototypicality, Spanish language development, syntax, dialect representation and variation, or the age of acquisition of skills or words.
Empty Bubble	Does not meet full or half bubble criteria.
Dash	Item development evidence was not provided.