Behavior Screening Tools Chart
Rating Rubric

*Please note that the following rubrics are applied separately for **each sub-scale, grade level/span, and informant** targeted by the tool.*

Technical Standard 1. Classification Accuracy

Note: Classification Accuracy will be rated separately for each criterion measure and time of year for the administration (e.g., Fall, Winter, Spring). Ratings will be provided for up to two different criterion measures and up to three different time points. Data for additional criterion measures or administration times may be reported but will not be rated.

Rating	Definition
Full Bubble	All of Q1 – Q3 rated as YES ***and*** (a) The lower bound of the confidence interval around the Area Under the Curve (AUC) estimate ≥ 0.75 ***or*** (b) If a confidence interval is not available, the lowest estimate of the AUC is ≥ 0.75* ***and*** Sensitivity ≥ 0.70 and Specificity ≥ 0.70
Half Bubble	All of Q1 – Q3 rated as YES ***and*** (a) The lower bound of the confidence interval around the AUC estimate is ≥ 0.70 but < 0.75 ***or*** (b) If a confidence interval is not available, the lowest estimate of the AUC is ≥ 0.70* ***or*** Sensitivity ≥ 0.60 and Specificity ≥ 0.60
Empty Bubble	Does not meet full or half bubble.

**Note: This option will only be included in the rubric for the 2017 and 2018 review cycles and will be phased out in 2019.*

Q1. Was an appropriate measure of social, emotional, or behavioral skills used as an outcome?

Q2. Was a convincing rationale provided for the selection of comparison point against which the screener was judged (e.g., percentile, cut score)?

Q3. 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 among 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 with at least two sources of variance ***or*** (b) At least two other 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 students across all performance levels ***and*** For each type of reliability reported, (a) the lower bound of the confidence interval around the median estimate met or exceeded 0.70 ***or*** (b) if a confidence interval is not available, the lowest estimate met or exceeded 0.70*
Half Bubble	(a) A model-based approach to reliability was reported with at least two sources of variance ***or*** (b) At least two other 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 students across all performance levels ***and/or*** For each type of reliability reported, (a) the lower bound of the confidence interval around the median estimate fell below 0.70 but met or exceeded 0.60 ***or*** (b) if a confidence interval is not available, the lowest estimate fell below 0.70 but met or exceeded 0.60*
Empty Bubble	Does not meet full or half bubble
Dash	Reliability data were not provided

**Note: This option will only be included in the rubric for the 2017 and 2018 review cycles and will be phased out in 2019.*

Technical Standard 3: Validity

Rating	Definition
Full Bubble	There are at least two types of appropriately justified validity analyses* from a sample representative of students across all performance levels ***and*** (a) the lower bound of the confidence interval around each standardized estimate met or exceeded 0.60 (or if not, was within an acceptable range given the expected relationship with the criterion measure(s)) ***or*** (b) if a confidence interval is not available, the lowest estimate exceeded 0.60**
Half Bubble	Analyses, measures, and sample were appropriate, but evidence was mixed, with the lower bound of the confidence interval for one or more, but not all, estimate(s) either not meeting or exceeding 0.60 or not within an acceptable range given the expected relationship with the criterion measure(s)
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.*
***Note: This option will only be included in the rubric for the 2017 and 2018 review cycles and will be phased out in 2019.*

Technical Standard 4: Sample Representativeness

Rating	Definition
Full Bubble	At least one classification accuracy analysis was conducted using a large representative national sample (at least 150 students across at least three geographic divisions*) ***and*** Cross-validation (i.e., multiple studies)
Half Bubble	At least one classification accuracy analysis was conducted using a large representative national sample (at least 150 students across at least three geographic divisions) or multiple regional/state samples with no cross-validation ***or*** One or more regional/state samples with cross-validation
Empty Bubble	Classification accuracy analysis was conducted using one regional or state sample with no cross-validation ***or*** One or more local samples

**Nine geographical divisions as defined by U.S. Census Bureau:*
https://www.census.gov/geo/reference/gtc/gtc_census_divreg.html

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	One or more of the following three types of analyses were conducted: 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) 4. Testing differential classification accuracy across demographic groups
No	Does not meet "yes"