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 screening tool.*

1. Classification Accuracy

Note: Classification Accuracy is rated separately for each criterion measure and time of year of administration (e.g., Fall, Winter, Spring). Ratings are provided for up to two 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 (see list below table) ***and*** The median estimate of the Area Under the Curve[1] (AUC) is ≥ 0.75 ***and*** Sensitivity ≥ 0.70 and Specificity ≥ 0.70.
Half Bubble	All of Q1 – Q3 rated as YES ***and*** (a) The median estimate of the AUC is ≥ 0.70 ***or*** (b) Sensitivity ≥ 0.60 and Specificity ≥ 0.60.
Empty Bubble	Does not meet full or half bubble.

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

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

Q3. Were the classification analyses and cut-points adequately performed?

[1] 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.

2. Reliability

Rating	Definition
Full Bubble	At least two types of reliability were reported that are appropriate[1] for the purpose of the tool, ***and*** the analyses are drawn from at least two samples that are representative of students across all performance levels, ***and*** the median of the estimates for each type met or exceeded 0.70.
Half Bubble	At least two types of reliability were reported that are appropriate[1] for the purpose of the tool, ***and*** (a) the analyses are drawn from one sample representative of students across all performance levels, and the median of the estimates met or exceeded 0.70 ***or*** (b) the analyses are drawn from at least two samples representative of students across all performance levels and the median of the estimates for each type met or exceeded 0.60.
Empty Bubble	Does not meet full or half bubble.
Dash	Reliability data were not provided.

[1] Tests which require human judgment must report inter-rater reliability to be eligible for a Full or Half Bubble rating. Other types of reliability must include justification of appropriateness given the purpose of the tool.

3. Validity

Rating	Definition
Full Bubble	At least two types of appropriately justified[1] validity analyses are reported, ***and*** the analyses are drawn from at least one sample representative of students across all performance levels, ***and*** the median of the estimates for each met or exceeded 0.60 (or was within an acceptable range given the expected relationship with the criterion measure(s)).
Half Bubble	One type of appropriately justified[1] validity analysis is reported, ***and*** the analysis is drawn from a sample representative of students across all performance levels, ***and*** the median of the estimates met or exceeded 0.60 (or was 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.

[1] Appropriately justified types of validity 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.

4. Sample Representativeness

Description	Definition
National with Cross-Validation	At least one classification accuracy analysis was conducted using a national sample[1], ***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[1], ***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.

[1]A national sample consists of at least 150 students across at least three of nine geographical divisions defined by U.S. Census Bureau: *https://www.census.gov/geo/reference/gtc/gtc_census_divreg.html*

5. Bias Analysis Conducted

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 Item Response Theory (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"