NCRTI Technical Review Committee on Screening Tools: Technical Standards Defined

Classification Accuracy

Classification accuracy refers to the extent to which a screening tool is able to accurately classify students into "at risk for reading disability" and "not at risk for reading disability" categories.

Technical Standard 1: Classification Accuracy	
Rating	**Rating defined**
Full bubble: Convincing Evidence	Area Under the Curve (AUC) > 0.90 and all of Q1 – Q4 rated as YES
Half bubble: Partially Convincing Evidence	0.80 < Area Under the Curve (AUC) < 0.90 or 1 of Q1 – Q4 rated as NO
Open bubble: Unconvincing Evidence	Area Under the Curve (AUC) < 0.80 or 1 of Q1 – Q4 rated as NO
Dash: No evidence provided	Evidence was not provided

Q1. Was an appropriate external nationally normed measure of reading used as an outcome?

Q2. Were the children in the study only involved in general classroom instruction (i.e., they were **not** involved in a specialized tutoring program)?

Q3. Was risk adequately defined within an RTI approach to screening (e.g., 20th %-tile)?

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 among students with satisfactory and unsatisfactory reading performance, whereas values at .50 indicate the predictor is no better than chance. AUC values above based on the following:

Swets, J.A., Dawes, R. M., Monahan, J. (2000). Psychological science can improve diagnostic decisions. *Psychological Science in the Public Interest, 1*(1), 1-26.

Swets, J.A. (1992). The science of choosing the right decision threshold in high-stakes diagnostics. *American Psychologist, 47*(4), 522-532.

Swets, J.A. (1988). Measuring the accuracy of diagnostic systems. *Science, 240*(4857), 1285-1293.

Swets, J.A. (1986). Form of empirical ROCs in discrimination and diagnostic tasks: Implications for theory and measurement of performance. *Psychological Bulletin, 99*(2), 181-198.

Generalizability

Generalizability refers to the extent to which results generated from one population can be applied to another population. A tool is considered more generalizable if studies have been conducted on larger, more representative samples.

Technical Standard 2: Generalizability	
Rating	**Rating defined**
Broad	Large representative national sample with cross-validation
Moderate High	Large representative national sample or multiple regional/state samples with no cross-validation OR One or more regional/state samples with cross-validation
Moderate Low	One regional/state sample with no cross-validation, or one or more local samples
Narrow	Convenience Sample
Dash: No evidence provided	Evidence was not provided

Reliability

Reliability refers to the consistency with which a tool classifies students from one administration to the next. A tool is considered reliable if it produces the same results when administering the test under different conditions, at different times, or using different forms of the test.

Technical Standard 3: Reliability	
Rating	**Rating defined**
Full bubble: Convincing evidence	Split-half, coefficient alpha, test-retest, or inter-rater reliability greater than 0.80 and Q1 rated as YES
Half Bubble: Partially convincing evidence	Split-half, coefficient alpha, test-retest, or inter-rater reliability greater than 0.60 but less than 0.80 and Q1 rated as YES
Empty bubble: Unconvincing evidence	Split-half, coefficient alpha, test-retest, or inter-rater reliability less than 0.60 or Q1 rated as NO
Dash: No evidence provided	Evidence was not provided

Q1. Was the type of reliability reported appropriate given the purpose of the tool? Y N

Validity

Validity refers to the extent to which a tool accurately measures the underlying construct that it is intended to measure.

Technical Standard 4: Validity

Rating	Rating defined
Full bubble: Convincing evidence	All of Q1 – Q3 rated as Yes
Half Bubble: Partially convincing evidence	1 of Q1 – Q3 rated as NO
Empty bubble: Unconvincing evidence	2 or 3 of Q1 – Q3 rated as NO
Dash: No evidence provided	Evidence was not provided

Q1. Was convincing evidence supporting content validity presented?

Q2. Was convincing construct validity presented (correlations above .70)?

Q3. Was convincing predictive validity presented (correlations above .70)?

Disaggregated Reliability, Validity, and Classification Data for Diverse Populations

Data are disaggregated when they are calculated and reported separately for specific sub-populations.

Technical Standard 5: Disaggregated Reliability, Validity, and Classification Data for Diverse Populations

Rating	Rating defined
Full bubble: Convincing evidence	At least two of the three types of data (classification, reliability, and validity) are disaggregated for at least 1 group AND meet the criteria for convincing or partially convincing.
Half Bubble: Partially convincing evidence	One of the three types of data is disaggregated for at least 1 group AND meets the criteria for convincing or partially convincing.
Empty bubble: Unconvincing evidence	One or more of the three types of data are disaggregated for at least 1 group, but all of the disaggregated data meet the criteria for unconvincing.
Dash: No evidence provided	None of the data are disaggregated for diverse populations.

This document was produced under U.S. Department of Education Cooperative Agreement #H326E070004 with the American Institutes for Research. Grace Zamora Durán and Tina Diamond served as the project officers. The views expressed herein do not necessarily represent the positions or policies of the Department of Education. No official endorsement by the U.S. Department of Education of any product, commodity, service or enterprise mentioned in this publication is intended or should be inferred.