INFORMATION BRIEF

# The Imagine Language & Literacy Benchmark Test: Evidence of Technical Rigor

## Overview of Technical Rigor of the Imagine Language & Literacy Benchmark Test

RMC Research conducted research to analyze the technical properties of Imagine Learning's Language & Literacy Benchmark Test. Analyses and ratings were guided by technical criteria used by the National Center on Intensive Intervention (NCII) to rate the rigor of academic screening tools.[1]

> The Imagine Language & Literacy Benchmark Test demonstrates technical rigor.

### Summary of Classification Accuracy, Reliability, and Validity Analysis Results

| Technical Criterion | K | Grade 1 | Grade 2 | Grade 3 | Grade 4 | Grade 5 | Grade 6 |
|---|---|---|---|---|---|---|---|
| Classification Accuracy | ✓✓ | ✓✓ | ✓✓✓ | ✓✓✓ | ✓✓ | ✓✓ | ✓✓ |
| Reliability | ✓✓✓ | ✓✓✓ | ✓✓✓ | ✓✓✓ | ✓✓✓ | ✓✓✓ | ✓✓✓ |
| Validity | ✓✓ | ✓✓✓ | ✓✓✓ | ✓✓✓ | ✓✓ | ✓✓ | ✓✓✓ |

✓ Low   ✓✓ Moderate   ✓✓✓ High

Additional details about the study and findings are presented below.

## What is the Imagine Language & Literacy Benchmark Test?

- The Imagine Language & Literacy Benchmark Test is a computerized, adaptive screening instrument for identifying students in grades K–6 who are at risk for not meeting expected outcomes in literacy development.
- The test is made up of subtests that measure letter recognition, phonemic awareness, word recognition, basic reading vocabulary, sentence cloze, beginning book comprehension, leveled book comprehension, and cloze.
- The test is administered at the beginning-, middle-, and end-of-the school year.
- Students generally take between 10 and 25 minutes to complete the test.

## The Study

- The study includes over 5,000 grade K–6 students in eight states (Arizona, California, Delaware, Georgia, Illinois, Louisiana, Ohio, and Wyoming) across four US regions. The number of students included in analytic samples vary across psychometric criteria evaluated (e.g., reliability and validity analyses).
- Students completed the Imagine Language & Literacy Benchmark and the Measures of Academic Progress (MAP) Reading Assessment between January and March 2017 and again between April and June 2017.

This research brief presents selected results from analyses conducted by Dr. Emma V. Espel, Dr. Stephen J. Meyer, and Dr. Shelley H. Billig of RMC Research in 2017. For more information, contact them at billig@rmcres.com or call 800-922-3636.

[1] https://intensiveintervention.org/resource/screening-tools-chart

## Findings

# Classification Accuracy: **Moderate to High**

Classification accuracy indicates how accurately a test identifies students who are at risk for poor academic outcomes.

| Three types of statistics used | |
|---|---|
| Area under the curve (AUC) | .76 to .90 |
| Sensitivity | .75 to .98 |
| Specificity | .71 to .89 |

**Area under the curve (AUC)** statistics indicate the extent to which the test correctly classfied students.

**Sensitivity** statistics reflect the proportion of students who were correctly classified by the assessment as "at risk."

**Specificity** statistics reflect the proportion of students who were correctly classified as "not at risk." Presented for 70% sensitivity.

**Classification accuracy was assessed using a Receiver Operating Characteristic (ROC) analysis** of students in four states who took the winter Imagine Language & Literacy Benchmark and the spring MAP Assessment.

# Reliability: **High**

Reliability indicates the consistency with which a screener classifies students across multiple administrations.

| Statistic used | |
|---|---|
| Separation reliability | .87 to .91 |

**Separation reliability** statistics are used to estimate the extent to which the test generates consistent results.

**Reliability was assessed using a Rasch model-based approach** with students in eight states who took the winter Imagine Language & Literacy Benchmark.

# Validity: **Moderate to High**

Validity indicates the extent to which the screener measures what it is intended to measure as indicated by its correlation to another similar measure.

| Three types of statistics used | |
|---|---|
| Concurrent validity (winter) | .67 to .83 |
| Concurrent validity (spring) | .51 to .75 |
| Predictive validity | .53 to .75 |

**Concurrent validity using winter scores** on the Imagine Language & Literacy Benchmark and MAP.

**Concurrent validity using spring scores** on the Imagine Language & Literacy Benchmark and MAP.

**Predictive validity** using winter scores on the Imagine Language & Literacy Benchmark and spring scores on the MAP.

**Validity was assessed using bivariate correlations** among scores of students in four states who took the winter and spring Imagine Language & Literacy Benchmark and MAP assessments.

# No Evidence of Consistent Bias

Few items differ across student groups.

| Items differed (out of 138 possible) | |
|---|---|
| Males vs. females | 0 items |
| ELL vs. non-ELL | 6 items |
| White vs. non-white | 3 items |

**Bias was assessed using a Rasch model-based approach** among K–6 students in eight states who took the winter Imagine Language & Literacy Benchmark Test. Differerential item functioning (DIF) statistics were used to assess item difficulty across groups of students according to **(1)** gender, **(2)** English language learner (ELL) status, and **(3)** race/ethnicity (white/non-white).

**Study results support continued confidence in the validity and reliability levels of the Imagine Language & Literacy Benchmark Test results.**