Psychometric Evidence of FastBridge Universal Screening & Progress Monitoring System

2021

Illuminate Education

Psychometric Evidence of FastBridge Universal Screening & Progress Monitoring System

Copyright © 2021 by Illuminate Education, Inc. All rights reserved.
Warning: No part of this publication may be reproduced or transmitted in any form or by any means, electronic or mechanical, now known or later developed, including, but not limited to, photocopying, recording, or the process of scanning and digitizing, transmitting, or storing in a database or retrieval system, without permission in writing from the copyright owner.

Published and distributed by Illuminate Education, Inc.
6531 Irvine Center Dr.
Suite 100
Irvine, CA 92618

Email: sales@fastbridge.org
Website www.illuminateed.com
Phone: 949-656-3133

Illuminate Education partners with K–12 educators to equip them with data to serve the whole child and reach new levels of student performance. Our solution brings together holistic data and collaborative tools and puts them in the hands of educators. Illuminate supports over 17 million students and 5,200 districts/schools.

© 2021 Illuminate Education. All rights reserved.

Table of Contents

Item & Test Bias 6

aReading 6

Overview 6

Standard Error of Measurement 6

Reliability 7

IRT-based Marginal Reliability 7

Test-Retest Reliability 8

Content Validity 9

Concurrent & Predictive Validity 9

Classification Accuracy 12

aMath 16

Overview 16

Standard Error of Measurement 16

Reliability 17

IRT-Based Marginal Reliability 17

Test-Retest Reliability 18

Content Validity 19

Concurrent & Predictive Validity 19

Classification Accuracy 20

CBMreading 24

Overview 24

Reliability 24

Alternate Form Reliability 24

Test-Retest Reliability 25

Content Validity 25

Concurrent & Predictive Validity 26

Classification Accuracy 27

SAEBRS & mySAEBRS 30

Overview 30

Reliability 31

SAEBRS 31

mySAEBRS 32

Construct Validity 33
SAEBRS 33
mySAEBRS 34
Criterion Validity & Classification Accuracy 36
SAEBRS 36
mySAEBRS 38
Quantitative Evaluation of Bias 39
Approach 39
Study 1 40
Study 2 40
FastBridge Screening Data Study 41
earlyReading 42
Overview 42
Reliability 43
Internal-Consistency Reliability 44
Alternate-Form Reliability 45
Inter-Rater Reliability 45
Test-Retest Reliability 46
Validity 47
Content Validity 47
Criterion Validity 47
Classification Accuracy 50
earlyMath 51
Overview 51
Reliability 52
Internal-Consistency Reliability 52
Alternate-Form Reliability 53
Test-Retest Reliability 54
Inter-Rater Reliability 55
Validity 56
Content Validity 56
Criterion Validity 56
Classification Accuracy 60
AUTOreading 61

Overview 61
Reliability 62
Validity 63
Content Validity 63
Construct Validity 63
CBMmath Automaticity 65
Overview 65
Reliability 65
Validity 66
References 68

Item & Test Bias

Illuminate Education is committed to providing psychometrically sound assessments that provide valid and fair results for all students. One component of this commitment is the evaluation of how our tests and test items function in specific groups based on gender and race/ethnicity. Test items and content that are embedded in contexts are prone to bias because the context may be more familiar to one group than another. This form of bias can be evaluated statistically at the item level using a method known as differential item functioning (DIF) and at the test level using methods such as differential prediction. In addition to statistical evaluation of existing items, we use best practices to limit the possibility of offensive or biased content during the development process. We follow the recommended AERA/APA/NCME Standards for Educational and Psychological Testing guidelines and the Smarter-Balanced Assessment Consortium (SBAC) guidelines for promoting diversity and evaluating fairness (AERA/APA/NCME, 2014). For a description of the methods and the results from our statistical evaluation of test and item bias of CBMreading, aReading, aMath, SAEBRS, and mySAEBRS, see the *Statistical Analysis of Bias of FastBridge Assessments* (Illuminate Education Inc., 2020).

Many FastBridge assessments measure skills directly, free from additional contexts. These include most of the earlyReading and earlyMath assessment suites as well as CBMmath Automaticity and CBMmath Process. An example from earlyReading is the Letter Names measure, which presents students with letters from the English alphabet on a page to evaluate letter-recognition skills. This direct format for assessing a skill is not prone to bias. The same can be said about CBMmath Automaticity and CBMmath Process. Both sets of measures use a context-free and direct assessment of math computational skills.

The process of promoting fairness and diversity in our assessments is ongoing. We are implementing new approaches to promote diversity in new content development, including ensuring a diverse author pool and expert bias reviews prior to publishing new items and passages. The landscape and expectations have changed, and it will take time to satisfy these important changes.

aReading

Overview

The FastBridge aReading assessment is a computer-adaptive test that assesses the full range of reading skills from kindergarten through grade 12. By employing an adaptive test algorithm, the aReading test accurately estimates each student's reading ability across the full range of abilities in each grade. It is intended as a universal screener to identify students at risk for academic delays and to provide results that help educators tailor instruction for each student's needs. For a detailed description of the development and content coverage of aReading, see the *FastBridge Assessments Content Description and Use Guide* (Illuminate Education, Inc. 2020).

Standard Error of Measurement

Table 1 reports the distribution of the conditional standard error of measurement (cSEM) for all students completing FastBridge aReading in a screening period. The data for K–3 were collected in the winter of 2018. The data for Grades 4–8 were collected in the fall of 2019.

The maximum likelihood estimation algorithm used to calculate aReading scores from the vector of responses (1s and 0s) for each student also produces a standard error value associated with each vector. The mean of the standard error averaged across a large representative sample of students is an index of the population error variance used in the computation of reliability in classical test theory. The cSEM generally forms a U-shape curve when plotted against the ability or scale score metric, indicating higher error rates among the lowest and highest scores. Thus, the cSEM provides a more accurate picture of the precision of a test than the SEM derived from classical reliability coefficients such as Cronbach's alpha.

The table below includes the mean and the 5th and 95th percentile of the distribution of cSEMs by grade. Note the aReading scale score has a mean of 500 and a standard deviation of 50. Relative to the standard deviation of the aReading scale, the results indicate that the adaptive algorithm results in very precise score measurement across the full range of abilities.

Table 1

aReading Conditional Standard Error of Measurement

Grade Level	Sample Size	5th	Mean	95th
K	16,076	5.84	7.92	10.48
1	52,350	4.71	6.06	8.14
2	125,435	4.49	5.34	6.91
3	122,701	4.43	5.10	5.94
4	124,400	4.50	5.11	5.78
5	123,765	4.55	5.09	5.71
6	111,058	4.56	5.12	5.83
7	89,714	4.57	5.18	6.15
8	85,698	4.59	5.29	6.66

Reliability

IRT-based Marginal Reliability

With adaptive Item Response Theory (IRT)-based assessments, reliability is a function of the student's IRT-based ability estimate and the conditional standard error of the ability estimate. The method mostly closely approximating the traditional internal consistency reliability method is referred to as empirical reliability.

$$r_{xx} = \frac{Var(\hat{\theta})}{Var(\hat{\theta}) + SE(\hat{\theta})^2}$$

Where r_{xx} is the reliability, $Var(\hat{\theta})$ is the total variance of the theta ability estimates in the population, and $SE(\hat{\theta})^2$ is the mean squared conditional standard error of the thetas across the population. Marginal reliability assumes a known ability distribution and item parameters (Dimitrov, 2003).

The table below describes the sample used for the computation of reliability and the overall reliability estimate with lower (LB) and upper bounds (UB) for a 95% confidence interval. These coefficients provide credible evidence of strong reliability of an aReading score in each grade.

Table 2

aReading Internal Consistency Reliability Estimates

Grade	Count	Reliability Coefficient	Lower Bound	Upper Bound	Female	White	Black	Hisp.	Asian	Other
K	45,437	0.91	0.90	0.91	53	63	9	15	7	4
1	69,394	0.95	0.94	0.95	51	64	11	11	6	6
2	106,647	0.96	0.95	0.96	51	63	10	11	7	9
3	108,008	0.94	0.94	0.95	51	63	11	12	7	7
4	124,400	0.95	0.95	0.95	51	63	11	13	7	6
5	123,765	0.95	0.95	0.95	52	63	11	13	6	7
6	111,058	0.95	0.95	0.95	52	63	12	13	6	7
7	89,714	0.96	0.96	0.96	52	62	13	13	6	7
8	85,698	0.96	0.96	0.96	52	62	12	13	6	8

Test-Retest Reliability

Test-retest reliability was computed on a sample of students who were retested on aReading within a 60-day window. Retesting occurs in a very small fraction of administrations in which the adaptive algorithm did result in a sufficiently precise score. In those instances, the student is presented with new items that were not present in the first administration. Because retesting only occurs when the initial administration did not converge on an accurate estimate of ability, these retest coefficients represent a lower-bound estimate of true retest reliability. Table 3 describes sample size and test-retest reliability coefficients for aReading.

Table 3

aReading Test-Retest Reliability by Grade

Grade	Count	Reliability Coefficient	Lower Bound	Upper Bound
K	534	0.74	0.70	0.78
1	1,134	0.85	0.83	0.87
2	1,360	0.90	0.89	0.91
3	1,639	0.90	0.89	0.91
4	2,009	0.93	0.92	0.94
5	1,548	0.92	0.91	0.93

6	943	0.90	0.89	0.91
7	645	0.80	0.77	0.83
8	529	0.86	0.84	0.88

Content Validity

Content validity is informed by the theory, content, and test development process. See the *FastBridge Assessments Content Description and Use Guide* (Illuminate Education, Inc. 2020) for a description of the underlying theory, the content blueprint, and the item development process.

Concurrent & Predictive Validity

Concurrent validity of FastBridge aReading tests was examined using the Gates MacGinitie Reading Tests-4th Edition (GMRT-4th; MacGinitie, MacGinitie, Maria, & Dreyer, 2000). The GMRT-4th is a norm-referenced, group-administered measure of reading achievement distributed by the Riverside Publishing Company. It is designed to provide guidance in planning instruction and intervention and is typically used as a diagnostic tool for general reading achievement. The GMRT-4th was normed with students in the pre-reading stages through high school levels.

The GMRT-4th was selected because of its strong criterion validity. Correlations between the GMRT-4th composite score and comprehension and vocabulary subtests of the Iowa Test of Basic Skills and GMRT-4th composite scores across grades is high (0.76 and 0.78 respectively; Morsy, Kieffer, & Snow, 2010). A similar pattern of results was observed between the GMRT-4th and subscales of the California Tests of Basic Skills (.84 and .81 respectively; Morsy et al., 2010). GMRT-4th scores also correlate highly with Comprehensive Tests of Basic Skills (CTBS) vocabulary, comprehension, and composite scores (.72, .79, and. 83 respectively; Morsy, et al., 2010). Further, the correlation between GMRT-4th composite scores and reading scores on the Basic Academic Skills Samples were strong as well (.79; Jenkins & Jewell, 1992).

Five trained FastBridge aReading project team data collectors administered the GMRT-4th during February 2011 at two separate schools. Students were administered the word decoding/vocabulary and comprehension subtests of the GMRT-4th during two separate testing sessions. Some students were administered the word decoding/vocabulary section first while other students were administered the comprehension subtest first. Participants included students in first through fifth grades. Three classrooms per grade at School A participated (n = 622); all students in first through fifth grades at School B participated (n = 760). See Table 4 for demographic information disaggregated by school.

Descriptive information for the GMRT-4th and FastBridge aReading scores are provided in Table 5 and correlations coefficients are presented in Table 6.

Overall, there appears to be a strong positive correlation between composite scores from the GMRT-4th and aReading scaled scores. There is some variability between grades, with coefficient values between .64 and .83. Subtests showed greater variability. Specifically, comprehension correlation coefficients ranged from .58 to .81. Content, construct, and predictive validity of aReading is summarized in Table 7.

Table 4

Demographic for Concurrent and Predictive Validity

Category	School A	School B
White	70%	69%
Black	6%	5%
Hispanic	9%	6%
Asian/Pacific Islander	13%	19%
American Indian/Alaskan Native	3%	1%
Free/Reduced Lunch	19%	14%
LEP	14%	14%
Special Education	11%	10%

Note: LEP = Limited English Proficiency. Percentages are rounded to whole numbers and therefore may not add to 100.

Table 5

Descriptive Information of the aReading Validity Sample

	Decoding			Vocabulary			Comprehension			Composite		
Grade	N	Mean	SD	N	Mean	SD	N	Mean	SD	N	Mean	SD
1	348	436	47	-	-	-	130	407	44	125	409	43
2	163	449	43	-	-	-	215	459	49	-	-	-
3	-	-	-	170	485	41	168	484	42	165	484	40
4	-	-	-	182	504	39	180	503	42	175	502	36
5	-	-	-	182	513	31	187	518	35	181	514	30
1–5	511	442.5	45	534	501	39	881	477	56	646	483	53

Note: N = sample size, SD = standard deviation.

Table 6

Correlation Coefficients between GMRT-4th and FastBridge aReading Scaled Score

Grade	Decoding	Vocabulary	Comprehension	Composite
1	.82 (131)	-	.73 (130)	.83 (125)
2	.68 (163)	-	.75 (215)	-
3	-	.79 (170)	.81 (168)	.84 (165)
4	-	.76 (182)	.72 (180)	.78 (175)
5	-	.65 (182)	.58 (187)	.64 (181)
1–5	.75 (348)	.74 (534)	.82 (881)	.86 (646)

Note: Sample size is denoted by ().

Table 7

Content, Construct, and Predictive Validity of aReading

				Coefficient (if applicable)	
Validity	**Grade**	**Test or Criterion**	**n (range)**	**Range**	**Median**
Content	K–5	Reading teachers/experts			
Content	K–3	Items administered by Theta Level	287		
Predictive	1–5	Gates-MacGinitie	125–215	0.64–0.84	0.78
	1		125	0.83	
	2*		215	0.75	
	3		165	0.84	
	4		175	0.78	
	5		181	0.64	
Construct	1–5	Curriculum-Based Measurement of Oral Reading Fluency	55–171	0.56–0.83	0.80
	1		55	0.83	
	2		171	0.81	
	3		108	0.74	

	4		114	0.80	
	5		103	0.56	
Construct	1–5	MAP	55–398	0.69–0.83	0.77
	1		55	0.69	
	2		302	0.83	
	3		391	0.83	
	4		398	0.77	
	5		376	0.73	

Note: Grade 2 Predictive validity for the GMRT-4th is based on the comprehension subtest, whereas all other grades are based on the overall composite score. Measures of Academic Progress (MAP)

Classification Accuracy

Classification accuracy is another component of validity evidence that is particularly important for screening for risk. Originally, aReading classification accuracy was established from a sample of 777 students in first through fifth grades from two suburban schools in the Midwest. More recently, classification accuracy has been examined through prediction of state test proficiency levels. Using ordinal regression analysis, FastBridge researchers identified cut scores associated with each state test proficiency level: California, Georgia, Illinois, Minnesota, Michigan, New York, and Wisconsin.

The original study sample was comprised of 116 students in first grade, 188 in second, 159 in third, 156 in fourth, and 158 in fifth grade. Gender of the sample was approximately 49% female and 51% male. Approximately 67% of students in the sample were White, 5% Black, 19% Asian/Pacific Islander, 5% Hispanic, 2% American Indian/Alaskan Native, and 2% unspecified. In addition, 10% of students were receiving special education services, and 10% of students were classified as having limited English language proficiency. Socioeconomic status information was not available for the sample, but the schools the students were drawn from had rates of free/reduced lunch of 13% and 23% in 2009–10. Cut scores for aReading to predict students "At Risk" and "Somewhat At Risk" for reading difficulties were developed for the Gates-MacGinitie Reading Tests–Fourth Edition (GMRT-4th; MacGinitie, et al., 2000) and the Measures of Academic Progress (MAP). Categories for the former were defined as students scoring below the 40th and 20th percentiles of the local sample, and cut scores for each category developed by an adjacent school district for MAP were used on this sample.

Table 8 shows the results from a Receiver Operating Characteristic (ROC) analysis including the Area Under the Curve (AUC), sensitivity, and specificity. The AUC coefficients are evidence of strong classification accuracy. The results also indicate that the cut scores provide strong sensitivity with compromising specificity.

Table 8

Classification Accuracy for aReading with MAP and GMRT-4th

MAP	GMRT-4th

Grade	Cut	AUC	Sens.	Spec.	Cut	AUC	Sens.	Spec.
	Some Risk (15th– 40th percentile)				**Some Risk (15th–40th percentile)**			
1	-	-	-	-	436	.91	.76	.86
2	490	.89	.77	.84	477	.87	.86	.71
3	527	.89	.89	.77	490	.89	.82	.87
4	537	.92	.82	.87	506	.82	.72	.82
5	547	.88	.93	.77	522	.85	.83	.76
	High Risk (<15th percentile)				**High Risk (<15th percentile)**			
1	-	-	-	-	430	.94	.88	.87
2	497	.89	1	.73	461	.88	.70	.93
3	517	.95	.95	.76	490	.92	.97	.77
4	537	.94	.96	.78	495	.94	.85	.92
5	537	.93	1	.82	506	.87	.85	.84

Table 9 summarizes the strength of the prediction of aReading Grade 3 scores with each state accountability test. The correlation coefficients represent concurrent and predictive validity. Most coefficients are above 0.70 and several in the 0.80s—thus demonstrating strong validity evidence. The area under the curve (AUC) results from the ROC curve analysis indicate that FastBridge aReading scores can be used to accurately identify students who are likely to pass the state test.

Table 9

aReading Correlation and AUC with State Accountability Tests, by Grade and Season

Grade	Season	State	N	Corr.	AUC
3	Fall	CA	1,834	0.75	0.89
	Winter	CA	1,864	0.77	0.91
	Spring	CA	1,838	0.77	0.92
4	Fall	CA	1,923	0.76	0.90
	Winter	CA	1,945	0.77	0.91
	Spring	CA	1,932	0.79	0.92
5	Fall	CA	1,904	0.75	0.90
	Winter	CA	1,953	0.78	0.91
	Spring	CA	1,931	0.79	0.92
6	Fall	CA	2,046	0.80	0.89
	Winter	CA	2,078	0.80	0.91

	Spring	CA	1,986	0.81	0.91
3	Fall	IL	1,412	0.75	0.87
	Winter	IL	1,438	0.73	0.86
	Spring	IL	1,443	0.78	0.90
4	Fall	IL	2,146	0.76	0.88
	Winter	IL	2,171	0.75	0.89
	Spring	IL	2,174	0.78	0.90
5	Fall	IL	2,211	0.73	0.86
	Winter	IL	2,244	0.76	0.87
	Spring	IL	2,254	0.76	0.90
6	Fall	IL	2,200	0.68	0.84
	Winter	IL	2,198	0.71	0.85
	Spring	IL	2,187	0.73	0.86
7	Fall	IL	3,018	0.73	0.85
	Winter	IL	3,014	0.73	0.85
	Spring	IL	3,032	0.76	0.87
8	Fall	IL	2,624	0.72	0.85
	Winter	IL	2,629	0.73	0.86
	Spring	IL	2,628	0.74	0.87
3	Fall	MI	4,958	0.66	0.83
	Winter	MI	5,058	0.69	0.84
	Spring	MI	5,094	0.73	0.89
4	Fall	MI	5,697	0.73	0.86
	Winter	MI	5,769	0.72	0.86
	Spring	MI	5,742	0.81	0.92
5	Fall	MI	6,185	0.72	0.86
	Winter	MI	6,194	0.74	0.88
	Spring	MI	6,311	0.80	0.92
6	Fall	MI	3,262	0.72	0.84
	Winter	MI	3,515	0.64	0.84
	Spring	MI	3,554	0.78	0.90
7	Fall	MI	3,511	0.77	0.88
	Winter	MI	3,633	0.77	0.89
	Spring	MI	3,616	0.80	0.90
8	Fall	MI	3,684	0.74	0.88
	Winter	MI	3,805	0.74	0.87
	Spring	MI	3,836	0.80	0.89
3	Fall	MN	3,592	0.81	0.91
	Winter	MN	2,726	0.82	0.91
	Spring	MN	2,799	0.84	0.92
4	Fall	MN	3,572	0.80	0.90
	Winter	MN	2,771	0.82	0.91
	Spring	MN	2,778	0.83	0.92
5	Fall	MN	3,414	0.78	0.92

	Winter	MN	2,473	0.83	0.92
	Spring	MN	2,571	0.85	0.93
	Fall	MN	2,492	0.78	0.90
6	Winter	MN	2,523	0.79	0.90
	Spring	MN	2,554	0.77	0.91
	Fall	MN	886	0.79	0.90
7	Winter	MN	878	0.82	0.91
	Spring	MN	398	0.83	0.95
	Fall	MN	873	0.74	0.88
8	Winter	MN	678	0.78	0.89
	Spring	MN	386	0.83	0.95
	Fall	NY	122	0.63	0.82
3	Winter	NY	126	0.76	0.84
	Spring	NY	126	0.81	0.88
	Fall	NY	138	0.75	0.90
4	Winter	NY	141	0.77	0.91
	Spring	NY	142	0.78	0.93
	Fall	NY	108	0.77	0.87
5	Winter	NY	109	0.77	0.86
	Spring	NY	109	0.80	0.88
	Fall	NY	-	-	-
6	Winter	NY	-	-	-
	Spring	NY	-	-	-
	Fall	NY	124	0.77	0.89
7	Winter	NY	126	0.71	0.87
	Spring	NY	126	0.75	0.87
	Fall	NY	117	0.73	0.91
8	Winter	NY	121	0.68	0.86
	Spring	NY	121	0.71	0.88
	Fall	WI	3,288	0.77	0.90
3	Winter	WI	-	-	-
	Spring	WI	3,313	0.79	0.91
	Fall	WI	3,882	0.73	0.88
4	Winter	WI	-	-	-
	Spring	WI	3,947	0.76	0.91
	Fall	WI	3,693	0.75	0.86
5	Winter	WI	-	-	-
	Spring	WI	3,753	0.77	0.89
	Fall	WI	3,368	0.76	0.89
6	Winter	WI	-	-	-
	Spring	WI	3,451	0.79	0.90
	Fall	WI	5,821	0.71	0.84
7	Winter	WI	-	-	-
	Spring	WI	5,870	0.73	0.87

	Fall	WI	5,321	0.66	0.85
8	Winter	WI	-	-	-
	Spring	WI	5,351	0.68	0.86
	Fall	GA	12,953	0.78	0.92
3	Winter	GA	13,473	0.80	0.94
	Spring	GA	12,267	0.80	0.94
	Fall	GA	12,076	0.78	0.93
4	Winter	GA	12,413	0.81	0.94
	Spring	GA	11,007	0.80	0.94
	Fall	GA	13,595	0.81	0.94
5	Winter	GA	14,232	0.81	0.94
	Spring	GA	12,208	0.81	0.94
	Fall	GA	11,938	0.79	0.94
6	Winter	GA	10,987	0.79	0.93
	Spring	GA	8,597	0.80	0.94
	Fall	GA	11,930	0.79	0.93
7	Winter	GA	10,711	0.77	0.93
	Spring	GA	8,034	0.79	0.93
	Fall	GA	13,087	0.79	0.94
8	Winter	GA	11,577	0.79	0.93
	Spring	GA	8,683	0.77	0.93

aMath

Overview

The FastBridge aMath assessment is a computer-adaptive test that assesses the full range of reading skills from kindergarten through grade 12. By employing an adaptive test algorithm, the aMath test accurately estimates each student's math ability across the full range of abilities in each grade. It is intended as a universal screener to identify students at risk for academic delays and to provide results that help educators tailor instruction for each students' needs. For a detailed description of the development and content coverage of aMath, see the *FastBridge Assessments Content Description and Use Guide* (Illuminate Education, Inc. 2020).

Standard Error of Measurement

Table 10 reports the distribution of the cSEM for all students completing FastBridge aMath in a screening period. The data for K–3 were collected in the winter of 2018. The data for grades 4–8 were collected in the fall of 2019.

The maximum likelihood estimation algorithm used to calculate aMath scores from the vector of responses (1s and 0s) for each student also produces a standard error value associated with each vector. The mean of the standard error averaged across a large representative sample of students is an index of the population error variance used in the computation of reliability in classical test theory. The cSEM generally forms a U-shape curve when plotted against the ability or scale score metric and indicates higher

error rates among the lowest and highest scores. Thus, the cSEM provides a more accurate picture of the precision of a test than the SEM derived from classical reliability coefficients such as Cronbach's alpha.

The table below includes the mean and the 5th and 95th percentile of the distribution of cSEMs by grade. Note the aMath scale score has a mean of 200 and a standard deviation of 15. Relative to the standard deviation of the aMath scale, the results indicate that the adaptive algorithm results in a very precise score measurement across the full range of abilities.

Table 10

aMath Conditional Standard Error of Measurement

Grade Level	Sample Size	5th	Mean	95th
K	38,580	1.54	2.68	4.85
1	63,372	1.23	1.95	2.86
2	93,542	1.16	1.62	1.82
3	96,345	1.18	1.53	1.71
4	72,785	1.19	1.48	1.81
5	74,842	1.20	1.58	1.92
6	55,940	1.20	1.61	1.96
7	41,974	1.29	1.73	2.14
8	39,199	1.33	1.75	2.14

Reliability

IRT-Based Marginal Reliability

With adaptive IRT-based assessments, reliability is a function of the student's IRT-based ability estimate and the conditional standard error of the ability estimate. The method mostly closely approximating the traditional internal consistency reliability method is referred to as empirical reliability.

$$r_{xx} = \frac{Var(\hat{\theta})}{Var(\hat{\theta}) + SE(\hat{\theta})^2}$$

Where r_{xx} is the reliability, $Var(\hat{\theta})$ is the total variance of the theta ability estimates in the population, and $SE(\hat{\theta})^2$ is the mean squared conditional standard error of the thetas across the population. Marginal reliability assumes a known ability distribution and item parameters (Dimitrov, 2003).

The table below describes the sample used for the computation of reliability and the overall reliability estimate with lower (LB) and upper bounds (UB) for a 95% confidence interval. These coefficients provide credible evidence of strong reliability of the aReading score in each grade.

Table 11

aMath Internal Consistency Reliability Estimates

Grade	Count	Rxx	LB	UB	Female	White	Black	Hisp.	Asian	Other
K	38,812	0.85	0.84	0.85	49	80	6	6	3	5
1	63,974	0.91	0.90	0.91	49	74	7	8	4	7
2	94,377	0.92	0.92	0.92	49	71	9	9	5	6
3	97,275	0.94	0.93	0.94	49	70	8	10	5	7
4	72,785	0.97	0.97	0.97	48	71	7	10	5	7
5	74,842	0.98	0.98	0.98	48	73	7	9	4	7
6	55,940	0.98	0.98	0.98	48	70	7	10	5	8
7	41,974	0.98	0.98	0.98	48	68	8	10	5	9
8	39,199	0.98	0.98	0.98	48	62	10	17	5	6

Test-Retest Reliability

Test-retest reliability was computed on a sample of students who were retested on aMath within a 60-day window. Retesting occurs in a very small fraction of administrations in which the adaptive algorithm did result in a sufficiently precise score. In those instances, the student was presented new items that were not present in the first administration. Because retesting only occurs when the initial administration did not converge on an accurate estimate of ability, these retest coefficients represent a lower-bound estimate of true retest reliability.

Table 12

aMath Test-Retest Reliability by Grade

Grade	Count	Reliability Coefficient	Lower Bound	Upper Bound
K	271	0.77	0.72	0.81
1	678	0.83	0.81	0.85
2	827	0.89	0.87	0.90
3	936	0.89	0.88	0.90
4	1,732	0.91	0.90	0.92
5	1,197	0.93	0.92	0.94
6	780	0.95	0.94	0.96
7	613	0.93	0.92	0.94
8	525	0.94	0.93	0.95

Content Validity

Content validity is informed by the test development process and content coverage. See the FastBridge Universal Screener Development Guide for a description of the underlying theory, the content blueprint, and the item development process.

Concurrent & Predictive Validity

Concurrent validity evidence is based on a sample of 432 students in kindergarten through fifth grade who completed FastBridge aMath and NWEA MAP. The MAP is a computer-adaptive test for students in grades two through five. The Group Mathematics Assessment and Diagnostic Evaluation (GMADE Williams, 2004) is a comprehensive, norm-referenced assessment of mathematical skills. Students complete the GMADE using paper and pencils. In the sample, 89 students were in kindergarten, 77 in first grade, 91 in second grade, 89 in third grade, 74 in fourth grade, and 76 in fifth grade. Gender of the sample was approximately 49% male and 51% female. Approximately 88% of the students in the sample were White, 6% Black, 3% Hispanic, 2% Asian/Pacific Islander, and 1% American Indian/Alaskan Native. In addition, approximately 8% of students were receiving free/reduced lunch and 15% were receiving special education services. See Table 13 for a sample description and Table 14 for concurrent validity results.

Predictive validity of FastBridge aMath was examined using the MAP and the GMADE. This study was based on the same sample used for the concurrent validation. The results, presented in Table 15, provide evidence for strong predictive validity.

Table 13

Validity and classification accuracy study samples

Category	Total	K	Grade 1	Grade 2	Grade 3	Grade 4	Grade 5
N	496	89	77	91	89	74	76
Gender (Male)	49%	40%	47%	48%	51%	53%	54%
White	88%	89%	90%	89%	85%	88%	90%
Black	6%	6%	3%	7%	6%	7%	8%
Hispanic	3%	2%	3%	2%	5%	3%	3%
Asian	2%	1%	4%	1%	5%	0%	0%
American Indian	1%	0%	0%	1%	0%	2%	0%
Special Education	15%	5%	10%	11%	24%	22%	18%
Free/Reduced Lunch	8%	2%	1%	0%	20%	18%	11%

Table 14

Concurrent Validity of aMath with NWEA MAP by Grade

		aMath	**MAP**		**GMADE**	
Grade	**N**	**M (SD)**	**M (SD)**	**Corr.**	**M (SD)**	**Corr.**
K	89	189.15 (6.07)	167.11 (13.64)	0.76	106.01 (15.25)	0.62
1	77	197.62 (5.91)	186.44 (7.88)	0.71	101.11 (14.00)	0.66

2	91	205.71 (6.57)	198.16 (11.02)	0.81	101.82 (16.58)	0.70
3	89	210.99 (6.81)	212.42 (10.72)	0.76	109.13 (12.80)	0.76
4	74	217.88 (11.60)	223.41 (14.33)	0.84	104.92 (15.20)	0.67
5	76	223.86 (11.80)	229.25 (14.25)	0.88	106.05 (16.24)	0.84

Table 15

Predictive Validity Coefficients for aMath with MAP and GMADE

Grade	N	MAP	N	GMADE
K	89	0.76	81	0.62
1	77	0.71	72	0.66
2	91	0.81	67	0.67
3	89	0.76	86	0.76
4	74	0.84	60	0.67
5	76	0.88	65	0.84

Classification Accuracy

Originally, aMath classification accuracy was established from a sample of 432 students in kindergarten through fifth grades from two suburban schools in the Midwest. More recently, classification accuracy has been examined through the prediction of state test proficiency levels. Using ordinal regression analysis, FastBridge researchers identified cut scores associated with each state test proficiency level: California, Georgia, Illinois, Minnesota, Michigan, New York, and Wisconsin.

The original classification accuracy study sample is the same NWEA MAP and GMADE data used in the predictive validity study. The table below show the sensitivity, specificity, and AUC using aMath to identify moderate and high-risk performance on MAP and GMADE.

Table 16

Classification Accuracy for aMath with MAP and GMADE

	MAP				**GMADE**			
Grade	**Cut**	**AUC**	**Sens.**	**Spec.**	**Cut**	**AUC**	**Sens.**	**Spec.**

	Some Risk (15th–40th percentile)				**Some Risk (15th–40th percentile)**			
K	188	.91	.84	.91	187	.81	.72	.80
1	193	.93	.84	1.00	196	.88	.79	.90
2	196	.98	.97	1.00	206	.82	.80	.92
3	208	.91	.77	1.00	207	.89	.96	1.00
4	206	.98	.91	1.00	215	.74	.67	.80
5	220	.94	.86	1.00	223	.97	.85	.88
	High Risk (<15th percentile)				**High Risk (<15th percentile)**			
					187	.75	.72	.75
					190	.83	.79	.83
					202	.92	.80	.80
					201	.98	.96	1.00
					206	.88	.95	.75
					218	.88	.84	.80

Table 17 summarizes the strength of the prediction of aMath grade 3 scores with each state accountability test. The correlation coefficients represent concurrent and predictive validity. Most coefficients are above 0.70 and several in the 0.80s—thus demonstrating strong validity evidence. The AUC results from the ROC curve analysis indicate that FastBridge aMath scores can be used to accurately identify students who are likely to pass the state test.

Table 17

aMath Correlation and AUC with State Accountability Tests, by Grade and Season

Grade	Season	State	N	Corr.	AUC
3	Fall	IL	1,416	0.72	0.88
	Winter	IL	1,454	0.80	0.87
	Spring	IL	1,457	0.83	0.90
4	Fall	IL	2,154	0.77	0.89
	Winter	IL	2,189	0.80	0.90
	Spring	IL	2,191	0.84	0.91
5	Fall	IL	2,209	0.79	0.90
	Winter	IL	2,253	0.80	0.90
	Spring	IL	2,267	0.83	0.95

	Fall	IL	2,191	0.78	0.91
6	Winter	IL	2,200	0.83	0.93
	Spring	IL	2,188	0.86	0.95
	Fall	IL	3,018	0.83	0.90
7	Winter	IL	3,030	0.85	0.91
	Spring	IL	3,027	0.86	0.94
	Fall	IL	2,624	0.78	0.91
8	Winter	IL	2,633	0.81	0.94
	Spring	IL	2,619	0.82	0.94
	Fall	MI	4,940	0.70	0.89
3	Winter	MI	5,067	0.79	0.91
	Spring	MI	5,067	0.86	0.94
	Fall	MI	5,704	0.73	0.89
4	Winter	MI	5,776	0.77	0.89
	Spring	MI	5,742	0.88	0.95
	Fall	MI	6,184	0.79	0.91
5	Winter	MI	6,202	0.85	0.93
	Spring	MI	6,337	0.86	0.94
	Fall	MI	3,366	0.84	0.91
6	Winter	MI	3,511	0.84	0.92
	Spring	MI	3,516	0.88	0.94
	Fall	MI	3,508	0.85	0.92
7	Winter	MI	3,642	0.86	0.94
	Spring	MI	3,567	0.85	0.95
	Fall	MI	3,711	0.84	0.94
8	Winter	MI	3,803	0.82	0.93
	Spring	MI	3,745	0.83	0.93
	Fall	MN	3,385	0.75	0.89
3	Winter	MN	2,476	0.77	0.90
	Spring	MN	2,035	0.82	0.93
	Fall	MN	3,531	0.74	0.89
4	Winter	MN	2,756	0.77	0.90
	Spring	MN	2,253	0.89	0.95
	Fall	MN	3,393	0.79	0.89
5	Winter	MN	2,176	0.84	0.93
	Spring	MN	2,059	0.86	0.95
	Fall	MN	2,271	0.80	0.90
6	Winter	MN	1,936	0.87	0.94
	Spring	MN	1,945	0.88	0.95
	Fall	MN	641	0.81	0.90
7	Winter	MN	302	0.89	0.95
	Spring	MN	296	0.89	0.96
8	Fall	MN	298	0.84	0.95
	Winter	MN	-	-	-

	Spring	MN	300	0.85	0.95
	Fall	NY	126	0.77	0.86
3	Winter	NY	129	0.79	0.88
	Spring	NY	129	0.83	0.90
	Fall	NY	137	0.82	0.93
4	Winter	NY	140	0.88	0.95
	Spring	NY	142	0.90	0.97
	Fall	NY	106	0.78	0.90
5	Winter	NY	109	0.83	0.94
	Spring	NY	110	0.86	0.96
	Fall	NY	-	-	-
6	Winter	NY	-	-	-
	Spring	NY	-	-	-
	Fall	NY	122	0.82	0.92
7	Winter	NY	122	0.83	0.94
	Spring	NY	120	0.85	0.96
	Fall	NY	118	0.82	0.85
8	Winter	NY	118	0.81	0.90
	Spring	NY	116	0.83	0.90
	Fall	WI	3,265	0.71	0.89
3	Winter	WI	-	-	-
	Spring	WI	3,304	0.74	0.90
	Fall	WI	3,882	0.60	0.85
4	Winter	WI	-	-	-
	Spring	WI	3,938	0.74	0.90
	Fall	WI	3,645	0.76	0.89
5	Winter	WI	-	-	-
	Spring	WI	3,750	0.80	0.93
	Fall	WI	3,378	0.77	0.86
6	Winter	WI	-	-	-
	Spring	WI	3,442	0.83	0.90
	Fall	WI	5,781	0.71	0.89
7	Winter	WI	-	-	-
	Spring	WI	5,825	0.80	0.91
	Fall	WI	5,346	0.77	0.89
8	Winter	WI	-	-	-
	Spring	WI	5,354	0.77	0.92
	Fall	GA	11,487	0.83	0.93
3	Winter	GA	11,942	0.85	0.94
	Spring	GA	10,830	0.87	0.95
	Fall	GA	11,850	0.83	0.93
4	Winter	GA	12,376	0.86	0.94
	Spring	GA	11,115	0.88	0.94
5	Fall	GA	13,345	0.84	0.91

	Winter	GA	13,911	0.86	0.92
	Spring	GA	12,088	0.86	0.92
	Fall	GA	11,886	0.83	0.93
6	Winter	GA	10,855	0.85	0.93
	Spring	GA	8,603	0.87	0.94
	Fall	GA	11,785	0.84	0.93
7	Winter	GA	10,529	0.85	0.93
	Spring	GA	7,723	0.85	0.94
	Fall	GA	12,662	0.81	0.92
8	Winter	GA	10,870	0.82	0.92
	Spring	GA	8,104	0.81	0.92

CBMreading

Overview

FastBridge Curriculum-Based Measurement for Reading (CBMreading) is a version of curriculum-based measurement of oral reading fluency (CBM-R), which was originally developed by Deno and colleagues to index the level and rate of reading achievement (Deno, 1985; Shinn, 1989). CBMreading is a simple, efficient, evidence-based assessment used for universal screening in grades 1 through 8 and progress monitoring for grades 1–12. CBMreading uses easy, time-efficient assessment procedures to determine a student's general reading ability.

Reliability

Two types of reliability are reported for CBMreading, alternate-form and test-retest. Alternate-form reliability is especially pertinent for assessments with multiple forms. CBMreading uses three passages for screening and includes another 20 unique passages for weekly progress monitoring. Because alternate-form reliability accounts for both the random measurement error attributable to a given testing occasion *and* error due to content differences across forms, it represents are more stringent estimate of score stability.

Alternate-Form Reliability

Alternate-form reliability results reported in Table 18 were derived from fall 2019 screening. Each student read three passages. The median correlation between the three passages is reported as the alternate-form reliability. Because the student's total score is the median score from the three passages, and therefore uses information from all three passages, the actual reliability of the screening score can be computed from the Spearman-Brown formula: $nr_{xx}/(1+(n-1)r_{xx})$, where n is the number of measurement observations (n=3) and r_{xx} is the reliability based on one observation (i.e., alternate form). The table also reports the 95% confidence interval bounds (LB, UB) and the percentage of the sample by gender and race/ethnicity.

Table 18

Alternate-Form Reliability for CBMreading

Grade	Count	Reliability Coefficient	Lower Bound	Upper Bound	R_{xx}(scr)	Female	White	Black	Hisp.	Asian	Other

1	31,958	0.96	0.96	0.96	0.99	48	59	9	9	11	12
2	160,745	0.95	0.95	0.95	0.98	49	68	8	11	6	7
3	145,331	0.94	0.94	0.94	0.98	49	69	8	11	5	7
4	121,414	0.93	0.93	0.93	0.98	48	69	8	11	5	7
5	107,789	0.93	0.93	0.93	0.98	48	73	7	10	4	6
6	50,645	0.93	0.93	0.93	0.98	48	74	8	10	3	5
7	23,468	0.94	0.94	0.94	0.98	46	65	12	13	4	6
8	20,110	0.92	0.92	0.92	0.97	45	60	14	15	5	6
9	2,450	0.88	0.87	0.89	0.96						
10	2,033	0.89	0.88	0.90	0.96						
11	1,425	0.89	0.88	0.90	0.96						
12	1,047	0.89	0.88	0.90	0.96						

Test-Retest Reliability

Test-retest reliability was computed using progress monitoring data from the fall and winter of the 2018–2019 school year. In progress monitoring (PM), students are assessed weekly with a different CBMreading passage. For this analysis, and to maximize the sample size, scores from the first and second PM administrations were used. The median number of days between administrations was 7 in grades 1–6, and 11 in grades 7 and 8. Results are reported in the table below.

Table 19

CBMreading Test-Retest Reliability by Grade

Grade	Count	Reliability Coefficient	Lower Bound	Upper Bound
K	12,048	0.92	0.92	0.92
1	24,484	0.91	0.91	0.91
2	21,895	0.89	0.89	0.89
3	19,878	0.89	0.89	0.89
4	16,892	0.87	0.87	0.87
5	7,233	0.87	0.86	0.88
6	1,716	0.88	0.87	0.89
7	1,486	0.86	0.85	0.87
8	12,048	0.92	0.92	0.92

Content Validity

Content validity is informed by the test development process and content coverage. See the FastBridge Universal Screener Development Guide for a description of the underlying theory, the content blueprint, and the passage development process.

Concurrent & Predictive Validity

Predictive and concurrent criterion validity for grades 1–6 were originally evaluated using several well-established standardized reading assessments, including: the Test of Silent Reading Efficiency and Comprehension (TOSREC), MAP, AIMSweb Oral Reading Fluency, and DIBELS Next Oral Reading Fluency. The results are described below. Where applicable, the interval between the administration of FastBridge CBMreading and the criterion assessment is stated. Concurrent validity coefficients for CBMreading grade-level passages are provided in Table 20. Predictive validity coefficients for CBMreading grade-level passages are provided in Table 21. The data were obtained from students in three states: Georgia, Minnesota, and New York.

Table 20

Concurrent Validity Coefficients for CBMreading

Criterion	Grade	N	Correlation
AIMSweb	1	399	0.95
	2	425	0.97
	3	402	0.95
	4	445	0.96
	5	447	0.96
	6	229	0.95
DIBELS Next	1	399	0.95
	2	463	0.92
	3	483	0.96
	4	485	0.95
	5	503	0.95
	6	225	0.95
MAP	1	-	-
	2	237	0.81
	3	231	0.78
	4	233	0.73
	5	219	0.66
	6	212	0.69
TOSREC	1	218	0.86
	2	246	0.81
	3	233	0.81
	4	228	0.79
	5	244	0.81
	6	222	0.82

Note: N = sample size.

Table 21

Predictive Validity Coefficients for CBMreading

Criterion	Grade	N	Time Lapse in Weeks		Correlation
			Mean	SD	
AIMSweb	1	385	18.68	3.04	0.91
	2	413	18.56	2.44	0.93
	3	391	18.98	2.50	0.91
	4	427	19.00	2.32	0.94
	5	431	19.00	2.51	0.93
	6	220	17.45	0.86	0.94
DIBELS Next	1	425	35.57	2.02	0.82
	2	80	35.93	1.61	0.74
	3	76	35.79	1.47	0.91
	4	74	35.67	1.41	0.90
	5	85	35.67	1.41	0.93
	6	-	-	-	-
MAP	1	-	-	-	-
	2	240	35.23	1.42	0.76
	3	233	35.47	1.27	0.73
	4	235	35.23	0.88	0.69
	5	220	35.29	1.13	0.65
	6	212	35.06	0.96	0.71
TOSREC	1	44	35.57	2.02	0.47
	2	35	35.94	1.61	0.56
	3	33	35.79	1.48	0.69
	4	35	35.67	1.41	0.52
	6	18	35.06	0.96	0.87

Note: N = sample size; SD = standard deviation.

Classification Accuracy

Classification accuracy for CBMreading was initially determined based on a sample of 1,153 students in Minnesota. Data were collected during the 2012–2013 school year. The sample consisted of approximately 45% males and 55% females. Approximately 20% of the students involved were eligible for free/reduced lunch. Most students were White (52%). The remainder of the sample consisted of approximately 30% Hispanic, 12% Black, 4% Asian/Pacific Islander, and 1% American Indian/Alaska Native. Approximately 15% of students were receiving special education services. All participants were proficient in English. Results are shown in Table 22.

Table 22

Classification Accuracy by Grade Level for CBMreading Passages

Grade Level	N	Cut Score	AUC	Sensitivity	Specificity	Classification	Time Lag	Criterion
20th Percentile								
1	171	16.5	0.81	0.75	0.63	0.71	2– 4 mo.	TOSREC
2	206	42.5	0.93	0.88	0.87	0.88	2– 4 mo.	
3	188	75.5	0.89	0.84	0.83	0.84	2– 4 mo.	
4	181	108.5	0.87	0.78	0.82	0.79	2– 4 mo.	
5	202	107.5	0.90	0.84	0.79	0.83	2– 4 mo.	
6	205	118.5	0.90	0.92	0.72	0.88	2– 4 mo.	
1	171	17	0.77	0.63	0.82	0.74	4 mo.	MAP
2	206	57	0.82	0.63	0.85	0.76	4 mo.	
3	188	88	0.80	0.77	0.75	0.76	4 mo.	
4	181	113	0.88	0.82	0.81	0.81	4 mo.	
5	202	101	0.89	0.70	0.91	0.86	4 mo.	
6	205	126	0.89	0.83	0.82	0.82	4 mo.	
1	171	21	0.79	0.78	0.70	0.73	8 mo.	MAP
2	206	63	0.82	0.83	0.68	0.72	8 mo.	
3	188	67	0.77	0.51	0.88	0.75	8 mo.	
4	181	104	0.89	0.80	0.86	0.84	8 mo.	
5	202	97	0.89	0.74	0.92	0.88	8 mo.	
6	205	126	0.89	0.85	0.82	0.83	8 mo.	
1	171	16	0.80	0.66	0.81	0.76	~1 yr.	MAP
2	206	82	0.90	0.87	0.84	0.86	~1 yr.	
3	188	88	0.89	0.82	0.81	0.82	~1 yr.	
4	181	114	0.87	0.83	0.76	0.78	~1 yr.	
5	202	108	0.89	0.80	0.85	0.84	~1 yr.	
6	205	126	0.85	0.77	0.79	0.79	~1 yr.	
30th Percentile								
1	171	16.5	0.81	0.75	0.63	0.71	2– 4 mo.	TOSREC
2	206	44.5	0.94	0.91	0.83	0.89	2– 4 mo.	
3	188	79.5	0.88	0.83	0.83	0.83	2– 4 mo.	
4	181	117.5	0.83	0.73	0.79	0.75	2– 4 mo.	
5	202	115.5	0.87	0.79	0.77	0.79	2– 4 mo.	
6	205	135.5	0.88	0.76	0.82	0.78	2– 4 mo.	
1	171	31	0.78	0.84	0.57	0.74	4 mo.	MAP
2	206	82	0.83	0.81	0.73	0.78	4 mo.	

3	188	85	0.77	0.57	0.86	0.66	4 mo.	
4	181	128	0.82	0.84	0.59	0.74	4 mo.	
5	202	125	0.82	0.65	0.75	0.69	4 mo.	
6	205	144	0.82	0.76	0.75	0.75	4 mo.	
1	171	24	0.78	0.74	0.70	0.73	8 mo.	MAP
2	206	82	0.78	0.77	0.54	0.67	8 mo.	
3	188	98	0.80	0.82	0.65	0.76	8 mo.	
4	181	125	0.84	0.66	0.85	0.75	8 mo.	
5	202	128	0.86	0.80	0.73	0.76	8 mo.	
6	205	144	0.79	0.74	0.70	0.71	8 mo.	
1	171	22	0.83	0.76	0.77	0.76	~1 yr.	MAP
2	206	82	0.91	0.72	0.89	0.76	~1 yr.	
3	188	104	0.85	0.63	0.92	0.71	~1 yr.	
4	181	122	0.85	0.70	0.86	0.78	~1 yr.	
5	202	135	0.82	0.77	0.68	0.73	~1 yr.	
6	205	144	0.80	0.76	0.75	0.75	~1 yr.	

Note: N = sample size; AUC = area under the curve.

Further classification accuracy analyses were conducted using the Minnesota Comprehensive Assessment-III (MCA-III). Students who completed CBMreading and the MCA-III Reading assessment in the spring of grades 3, 4, and 5 were included in the analysis. Results are reported in Table 23. Separate results are provided for the high risk and some risk CBMreading cut scores and include the AUC derived from ROC curve analysis, sensitivity, and specificity. Based on the criterion established by the National Center on Intensive Intervention (NCII), CBMreading demonstrates strong classification accuracy evidence.

Table 23

Classification Accuracy for CBMreading and Minnesota Comprehensive Assessment-III

Grade	N	CBM-R M (SD)	MCA-III M (SD)	Corr.	Cut Score	AUC	Sensitivity	Specificity
				High Risk				
3	852	139 (40)	348 (20)	0.76	132	0.88	0.80	0.79
4	818	165 (39)	447 (15)	0.71	154	0.87	0.80	0.78
5	771	165 (40)	552 (16)	0.70	152	0.89	0.80	0.79
				Somewhat High Risk				
3	852	139 (40)	348 (20)	0.76	142	0.86	0.78	0.76
4	818	165 (39)	447 (15)	0.71	165	0.83	0.75	0.71
5	771	165 (40)	552 (16)	0.70	164	0.84	0.77	0.76

Note: N = sample size; M = Mean; SD = Standard Deviation; AUC = area under the curve.

Additional classification accuracy analyses were conducted using the NWEA MAP Reading assessment as the criterion measure. The analysis was based on students who were screened using CBMreading in the

fall, winter, or spring and completed NWEA MAP in the spring. Scores were obtained during the 2018–2019 school year and include students from seven Midwestern states. Results confirm the utility of CBMreading as a very brief and highly accurate measure for classifying students as at-risk for reading difficulties.

Table 24

CBMreading Correlation and Classification Accuracy with NWEA MAP Reading, by Grade and Season

Grade	Season	N	Corr.	AUC	Accuracy	Sensitivity	Specificity
1	Fall	181	0.63	0.78	0.65	0.61	0.79
	Winter	325	0.69	0.79	0.71	0.69	0.80
	Spring	277	0.71	0.83	0.70	0.66	0.83
2	Fall	977	0.78	0.90	0.82	0.82	0.83
	Winter	707	0.77	0.90	0.85	0.86	0.85
	Spring	598	0.76	0.91	0.82	0.82	0.85
3	Fall	1,690	0.76	0.90	0.82	0.80	0.82
	Winter	1,235	0.72	0.88	0.80	0.81	0.80
	Spring	1,070	0.69	0.86	0.76	0.79	0.75
4	Fall	1,243	0.73	0.88	0.80	0.81	0.80
	Winter	1,159	0.70	0.88	0.80	0.79	0.81
	Spring	1,177	0.69	0.85	0.73	0.81	0.71
5	Fall	1,345	0.67	0.84	0.74	0.81	0.72
	Winter	1,289	0.63	0.83	0.73	0.81	0.71
	Spring	1,275	0.60	0.85	0.74	0.78	0.73
6	Fall	743	0.76	0.91	0.83	0.83	0.82
	Winter	768	0.61	0.87	0.77	0.76	0.83
	Spring	383	0.72	0.89	0.82	0.82	0.82
7	Fall	345	0.71	0.80	0.59	0.78	0.55
	Winter	223	0.66	0.83	0.66	0.79	0.64
	Spring	235	0.63	0.79	0.66	0.80	0.62
8	Fall	180	0.75	0.91	0.79	0.79	0.80
	Winter	95	0.79	0.90	0.81	0.76	0.82
	Spring	62	0.70	0.92	0.81	0.82	0.80

SAEBRS & mySAEBRS

Overview

FastBridge provides two instruments for the evaluation of social emotional learning: a self-report rating scale called mySAEBRS and a teacher version of the rating scale called SAEBRS. SAEBRS stands for Social Academic and Emotional Behavior Rating Scale. The rating scales are comprised brief behavioral statements (e.g., “I like school”) or dispositions (e.g., “I am happy”) that are rated on a 0 (Never) to 4-point (Almost Always) scale. The instrument provides an overall composite score and three subscale

scores: Social Behavior, Academic Behavior, and Emotional Behavior. The student form is comprised of 20 items and the teacher form is comprised of 19 items. Some items are reverse scored so that higher scores represent more positive behaviors.

Reliability

SAEBRS

The internal consistency reliability of SAEBRS Total Behavior scale and subscales was evaluated using classical coefficient alpha and coefficient omega derived from confirmatory factor models and inter-rater reliability. Internal consistency reliability is pertinent to any measure that consists of multiple items measuring a single broad construct. However, coefficient alpha generally underestimates reliability for multidimensional traits like those measured by SAEBRS. A model-based coefficient like omega is not prone to the effects of multidimensionality because it is computed under a bifactor model, wherein all items are related to both the general behavior factor and one of three narrow factors. Omega (ω) coefficients represent the proportion of variance to all factors common to an item set of interest. Hierarchical omega (ω_H) coefficients represent the proportion of variance attributable to a factor after controlling for all other factors. These latter statistics are particularly informative when examining measures corresponding to bifactor structures as items are presumed to be multidimensional and driven by both general and specific factors.

Internal reliability was evaluated across two studies. The first study was conducted in four urban elementary schools (K–5) located in the Midwestern United States (Kilgus, et.al., 2016). All general education teachers in each school chose to participate in this study. The teachers screened all students in their classroom, resulting in a sample of 68 teacher participants and 1,243 students. The sample was characterized by a diverse student population regarding ethnicity, including: White (54.5%), Black (28.6%), Hispanic (5.3%), and multiracial (8.4%) students. The free/reduced-price lunch rate across the four schools was equal to 65.1%. Test-retest reliability was obtained from the same sample through Pearson correlation coefficients between screening periods: fall, winter, and spring.

Table 25

Omega and Test-Retest Reliability, SAEBRS

Type of Reliability	Informant	Grade	Coefficient
Internal (Omega)	Teacher	K–5	0.98
Internal (H omega)	Teacher	K–5	0.87
Test-retest F-W	Teacher	K–5	0.78
Test-retest W-S	Teacher	K–5	0.82
Test-retest F-S	Teacher	K–5	0.73

In a separate study, internal consistency alpha coefficients were computed (Kilgus, Eklund, von der Embse, Taylor, & Sims, 2016). This study was conducted with 567 elementary students (52.9% female) and 34

classroom teachers. The sample was characterized by a diverse student population, including: White (50.1%), Black (34.4%), Hispanic (11.3%), and multiracial (3.7%) students.

Table 26

Internal Consistency Reliability Coefficient Alpha for SAEBRS by Grade Level

Measure	Grade	N	Coefficient Alpha
Social	Elementary	243	0.89–0.94
Social	Middle	359	0.93
Social	High	488	0.89
Academic	Elementary	243	0.90–0.92
Academic	Middle	359	0.92
Academic	High	488	0.93
Emotional	Elementary	219	0.83
Emotional	Middle	359	0.77
Total	Elementary	243	0.93
Total	Middle	359	0.94
Total	High	488	0.93

Evidence of inter-rater reliability was obtained on a sample of 488 high school students. Each student was rated by two teachers who taught the student in one high school period each day.

Table 27

Inter-Rater Reliability, SAEBRS Total Behavior Scale

Measure	Grade	Correlation	N
Social	High School	0.41	488
Academic	High School	0.47	488
General	High School	0.48	488

mySAEBRS

Reliability of mySAEBRS' subscales and the Total Behavior scale was evaluated as part of a larger study examining the factor structure, factor structure invariance across age levels, and item parameters (von der Embse, et.al., 2016; Kilgus, et.al., 2020). The data included unique responses from 24,094 students between kindergarten and 12th grade during the 2016–2017 academic year. Males comprised 53% of the sample. The racial distribution of students based on available data from 70% of the schools was as

follows: 40.3% White, 16.7% African American, and 6.9% Hispanic. The sample consisted of students from Kindergarten through Grade 8 in the following percentages: K = 2.1%, first = 2.1%, second = 6.4%, third = 14.8%, fourth = 15.0%, fifth = 14.6%, sixth = 14.8%, seventh = 11.0%, eighth = 11.1%, ninth = 2.1%, 10th = 2.5%, 11th = 2.1%, and 12th = 1.3%. Within this sample, schools reported that 4.6% of students received special education services. Among the special education sample, 27% had a SAEBRS rating in the at-risk range.

Internal consistency reliability was estimated using the omega coefficients obtained from the best fitting factor model. Omega coefficients were chosen over other commonly evaluated internal consistency statistics (e.g., coefficient alpha) given (a) the limitations associated with these latter statistics (Sijtsma, 2009) and (b) the applicability of omega coefficients to multidimensional factor structures, including bifactor models (Reise, et al., 2013). Three types of omega coefficients were calculated: omega (ω), which represents the proportion of variance attributable to all factors (i.e., both global and specific) common to the target items; ωH, which represents the proportion of variance attributable to the global factor after controlling for all specific factors; and ωS, which represents the proportion of variance attributable to the specific factors after controlling for the global factor. Acceptable reliability was defined as coefficients greater than .70 for ω values (Salvia et al., 2010) and .50 for ωH and ωS values (Gignac & Watkins, 2013; Reise, et al., 2013).

The results for omega (ω) were as follows: Total Behavior ω = 0.828, Academic Behavior ω = 0.767, Emotional Behavior ω = 0.767, and Social Behavior ω = 0.725. These results indicate that the Total Behavior score is highly reliable and that the subscale scores are also very reliable. The result for ωH was 0.464 for Total Behavior, suggesting that after accounting for the subscales specific factors, the global factor accounted for about 50% of the variance. The results for ωS were 0.371 for Academic Behavior, 0.630 for Emotional Behavior, and 0.541 for Social Behavior, indicating that after accounting for the global factor, the specific factors still accounted for some variance.

Table 28

Internal Consistency Reliability (Omega) Coefficients for mySAEBRS Scales

Scale	Omega (ω)	Omega (ωH & ωS)
Total Behavior	0.83	0.46
Academic Behavior	0.77	0.37
Emotional Behavior	0.77	0.63
Social Behavior	0.72	0.54

Construct Validity

SAEBRS

An important form of validation evidence for behavioral instruments is the extent to which the underlying factor structure conforms to expectations. To examine this component of validity, a series of confirmatory factor analyses (CFA) was conducted to evaluate the extent to which various factor models fit SAEBRS data. The three models were: (a) unidimensional model, wherein each mySAEBRS item loads onto a single broad factor; (b) correlated-factor model, wherein each item loads onto its corresponding narrow factor and the narrow factors are permitted to covary with one another; and (c) bifactor model,

wherein each item loads onto both its corresponding narrow factor and the overall broad factor (von der Embse, et.al., 2016).

Model fit was evaluated through a series of fit statistics, including the chi-square goodness-of-fit test, Tucker–Lewis Index (TLI), Comparative Fit Index (CFI), root mean square error of approximation (RMSEA), and standardized root-mean square residual (SRMR). Observed fit statistic values were compared with the following cut-offs in evaluating model fit: chi-square goodness-of-fit test, $p > .05$; RMSEA ≤ .08; CFI/TLI ≥ .90; and SRMR ≤ .08 (Little, 2013).

Table 29 shows the results of model fit from each model. The best fitting model was the bifactor model. It displayed excellent fit by all fit statistics.

Table 29

Fit Statistics Results from Confirmatory Factor Analysis

Fit Statistic	Unidimensional	Correlated Factor	Bifactor
Chi-square	3,243	1,824	670
RMSEA	0.106	0.077	0.040
SRMR	0.157	0.111	0.520
CFI	0.924	0.961	0.990

In a recent study, a random sample of students from the national fall 2019 FastBridge database was selected to evaluate factor structure. Using anonymized (deidentified) data, a sample was selected based on race and gender to match the U.S. school population. A total of 10,000 students from 338 school districts across the country. The sample was 51% male, 58% White, 16% African American, 17% Hispanic, 5% Asian, and 4% other. All grades from 2–8 were evenly represented.

For this analysis, a bifactor model with a single broad factor and three subfactors corresponding to academic, social, and emotional behavior was fit. Within such models, the broad factor initially accounts for item covariance, while subfactors then account for residual item covariance. Results confirm findings from previous studies that indicate the presence of three distinct subscales and a total behavior composite each contributing unique variance to the data.

Table 30

Bifactor Model Fit Statistics, SAEBRS

Fit Statistic	Value
RMSEA	0.089
SRMR	0.047
CFI	0.919
TLI	0.896

mySAEBRS

An important form of validation evidence for behavioral instruments is the extent to which the underlying factor structure conforms to expectations. To examine this component of validity, a series of confirmatory factor analyses (CFA) were conducted to evaluate the extent to which various factor models fit mySAEBRS data. The four models were: (a) unidimensional model, wherein each mySAEBRS

item loads onto a single broad factor; (b) correlated-factor model, wherein each item loads onto its corresponding narrow factor and the narrow factors are permitted to covary with one another; (c) higher order model, wherein each item loads on a narrow factor and narrow factors load on a single broad factor; and (d) bifactor model, wherein each item loads onto both its corresponding narrow factor and the overall broad factor (Kilgus, et.al, 2020).

Model fit was evaluated through a series of fit statistics, including the chi-square goodness-of-fit test, Tucker–Lewis Index (TLI), Comparative Fit Index (CFI), root mean square error of approximation (RMSEA), and standardized root-mean square residual (SRMR). Observed fit statistic values were compared with the following cut-offs in evaluating model fit: chi-square goodness-of-fit test, $p > .05$; RMSEA ≤ .08; CFI/TLI ≥ .90 (Browne & Cudeck, 1993; Hu & Bentler, 1995; Markland, 2007); and SRMR ≤ .08 (Little, 2013).

Table 31 shows the results of model fit from each model. The best fitting model was the bifactor model. This model became the basis for subsequent research. Although model fit using three specific factors (subscales) did not meet expected levels for all fit statistics, it did with RMSEA and SRMR. This may be evidence for additional subfactors (subscales) or correlated errors; however, a content analysis of the behaviors strongly favors the three-factor approach.

Table 31

Fit Statistics Results from Confirmatory Factor Analysis

Fit Statistic	Unidimensional	Correlated Factor	Higher Order	Bifactor
Chi-square	58,127	47,655	47,655	9,801
RMSEA	0.119	0.109	0.109	0.083
CFI	0.641	0.705	0.705	0.847
TLI	0.598	0.665	0.665	0.806
SRMR	0.111	0.104	0.104	0.072

In a recent study, a random sample of students from the national fall 2019 FastBridge database was selected to evaluate factor structure. Using anonymized (deidentified) data, a sample was selected based on race and gender to match the U.S. school population. A total of 10,000 students from 338 school districts across the country. The sample was 51% male, 58% White, 16% African American, 17% Hispanic, 5% Asian, and 4% other. All grades from 2–8 were evenly represented.

For this analysis, a bifactor model with a single broad factor and three subfactors corresponding to academic, social, and emotional behavior was fit. Within such models, the broad factor initially accounts for item covariance, while subfactors then account for residual item covariance. Criteria for evaluating minimally acceptable model fit were established a priori: RMSEA values ≤ 0.08 and CFI values ≥ 0.90.

Results confirm findings from previous studies that indicate the presence of three distinct subscales and a total behavior composite each contributing unique variance to the data. RMSEA met the criteria for good fit; however, the CFI was modestly lower than desired and may result from a couple of items with relatively low correlations.

Table 32

Bifactor Model Fit Statistics, mySAEBRS

Fit Statistic	Value

RMSEA	0.063
SRMR	0.059
CFI	0.840
TLI	0.800

Criterion Validity & Classification Accuracy

SAEBRS

Three studies were conducted to evaluate the validity of SAEBRS.

The first study completed during the 2010–2011 academic year used the Social Skills Improvement System (SSIS; Gresham & Elliott, 2008) as the criterion measure. The SSIS is a comprehensive teacher rating scale (83 items) used to assess the three domains of student behavior: social skills, problem behaviors, and academic competence. The SSIS was considered a particularly important criterion given the scale's assessment of social and academic functioning as well as its pertinence to problem behaviors, which are also assessed within each SAEBRS subscale. For this study, only the social behavior (6 items) and academic behavior (6 items) scales of SAEBRS were used.

This study was conducted across three public schools within a single school district in the Southeastern United States. In this district, 4% of students are English language learners, and 40% qualified for free/reduced lunch. The race/ethnicity composition of the sample was: 50.6% White, 32.5% African American, 10.3% Hispanic, 2.1% Asian, and 3.7% other (Kilgus, Chafouleas, & Riley-Tillman 2013).

Across the three schools, 56 K–5 teachers agreed to participate. Using a random number generator, researchers randomly selected five students for participation in each classroom, resulting in the identification of 276 student participants (four teachers requested to rate only four randomly selected students because of time constraints).

In the second study, the Behavior Assessment System for Children, 2nd Edition Behavioral and Emotional Screening System (BASC-2 BESS) was used as the criterion measure. The BASC-2 BESS is an established, reliable, and valid measure of maladaptive behaviors. The results were originally reported in Kilgus, Eklund, von der Embse, Taylor, & Sims (2016).

Participants included 34 elementary teachers and their 567 students (52.9% female). The race/ethnicity composition of the sample was: 50.1% White, 34.4% African American, 11.3% Hispanic, 0.5% Asian, and 3.7% multiracial. Overall, 61.9% of students qualified for free/reduced-priced lunch.

In the third study, the SRSS and the Student Inventory of Behavior Support (SIBS) were used as criterion measures (Kilgus, Sims, von der Embse, & Taylor, 2016). Participants included 17 teachers and 346 students. This sample comprised all the teachers and students from a single, rural Midwestern elementary school enrolling third-, fourth-, and fifth-grade students (teacher and student participation rate = 100%). The student body in this school was characterized by a relatively even split in gender, homogeneous ethnicity profile (i.e., 95% White), a 20% free/reduced price lunch rate, and low English language learner enrollment. The sample also included students with special needs.

Table 33 summarizes the criterion validity coefficients with each criterion assessment. All coefficients are moderate to high, indicating strong criterion validity. Among these criterion assessments, the BASC-

2 BESS is perhaps the most similar in purpose and design. The coefficient is very high (-0.93). The negative value is expected since higher scores on the BASC-2 BESS represent more maladaptive behavior whereas high scores on SAEBRS represent more positive behavior.

Table 33

Criterion Validity Coefficients of SAEBRS with Criterion Measures

Type of Validity	Informant	Grade	Test or Criterion	Sample Size	Median Coefficient
Concurrent	Teacher	K–5	SSIS-Social Skills	276	0.88
Concurrent	Teacher	K–5	SSIS-Problem Behaviors	276	-0.89
Concurrent	Teacher	K–5	SSIS-Academic Competence	276	0.61
Concurrent	Teacher	K–5	BESS	567	-0.93
Concurrent	Teacher	K–5	SRSS	346	-0.84
Concurrent	Teacher	K–5	SIBS	346	-0.67

ROC curve analysis has been used across multiple studies to identify SAEBRS cut scores (Kilgus, et al., 2013; Kilgus, Eklund, et al., 2016). SAEBRS cut scores are used in applied decision-making to differentiate between students who are at risk or not at risk for behavioral and emotional problems. Cut scores were selected using a linear algorithm, which prioritized sensitivity greater than or equal to .80 and specificity at or above .70 (Kilgus, et. al. 2013). The overall classification accuracy of each SAEBRS scale has also been evaluated using AUC statistics. AUC statistics correspond to the likelihood that a randomly selected at-risk student will have a lower SAEBRS score than a randomly selected not at-risk student. Researchers have suggested that AUC values between .50 and .70 be considered low, .70 to .90 moderate, and .90 to 1.00 high (Streiner & Cairney, 2007).

Classification accuracy was evaluated using the same data sets described above for the SSIS and the BASC-2 BESS. Table 34 and Table 35 show the classification accuracy results. Results provide solid evidence for the utility of SAEBRS as a screener for risk for social-emotional learning problems. Sensitivity, specificity, and AUC values all represent acceptable levels of accuracy.

Table 34

Classification Accuracy Statistics for SAEBRS and SSIS

Grade	Measure	N	Cut Score	Sensitivity	Specificity	AUC
Elementary	Social	243	12	0.87	0.83	0.93

Elementary	Academic	243	9	0.84	0.84	0.90
Elementary	Total	243	23	0.93	0.94	0.96

Note: N = sample size; AUC = area under the curve.

Table 35

Classification Accuracy Statistics for SAEBRS and BESS

Grade	Measure	N	Cut Score	Sensitivity	Specificity	AUC
Elementary	Social	219	12	0.87	0.83	0.93
Elementary	Academic	219	9	0.84	0.84	0.90
Elementary	Emotional	219	17	0.93	0.94	0.96
Elementary	Total	219	36	0.88	0.75	0.88
Middle	Social	359	12	0.87	0.86	0.94
Middle	Academic	359	9	0.90	0.77	0.94
Middle	Emotional	359	17	0.89	0.70	0.88
Middle	Total	359	36	0.94	0.91	0.98

Note: N = sample size; AUC = area under the curve.

mySAEBRS

A large study was conducted to evaluate the technical adequacy and diagnostic accuracy of mySAEBRS. The findings were originally published in von der Embse, Kilgus, Iaccarino, & Levi-Nelson (2017).

The study consisted of 1,102 students who attended two middle schools in an urban school district located in the Northeastern United States. Among the sample, 828 participants completed the mySAEBRS in January of 2016 and 274 participants completed the SDQ in May of 2016. The race/ethnicity composition of the sample was: 30.6% White, 31.3% African American, 7.8% Hispanic, 5.2% American Indian/Alaskan Native, 9.3% Asian American/Pacific Islander, and 15.8% as other. The sample was evenly divided between males and females. For the January administration, 34.5% of students were in grade 6, 30% in grade 7, and 35.5% in grade 8. For the May administration, 31.5% of students were in grade 6, 33.5% in grade 7, and 35% in grade 8.

The criterion measure was the Strength and Difficulties Questionnaire (SDQ; student version). The SDQ is a 25-item inventory that assess the social-emotional and behavioral health of children from the age of 4–17 (Goodman, 1997). Prior research suggests the SDQ consists of five distinct factors: emotion, conduct, hyperactivity, peer relations, and prosocial behavior, all of which additively inform the total SDQ score. Higher scores on emotion, conduct, hyperactivity, peer relations, and total SDQ score

indicate poor behavioral functioning; lower prosocial behavior scores are indicative of poor behavioral functioning.

For the current study, mySAEBRS Total Behavior predicted the SDQ total score; mySAEBRS Social scale predicted SDQ Emotion scale; and mySAEBRS Academic and Emotional scales predicted SDQ Peer scale. Table 36 provides internal consistency reliability coefficients for this sample as well as the bivariate correlation (validity coefficient) with SDQ and results from the ROC curve analysis. These results indicate that the mySAEBRS Total Behavior score is highly reliable and is moderately predictive of SDQ total scores. The AUC, sensitivity, and specificity based on the Total Behavior score met traditional criterion for acceptable classification accuracy. Overall, the mySAEBRS Total Behavior score correctly classified 79% of the students as at risk on the SDQ total score.

Table 36

mySAEBRS Reliability, Validity, and Classification Accuracy

Scale	Reliability	Validity	AUC (95% CI)	Cut Score	Sensitivity	Specificity	Accuracy
Social	0.64	-0.33	0.68 (.56–.80)	18	0.74	0.50	0.56
Academic	0.63	-0.31	0.74 (.62–.85)	13	0.74	0.64	0.65
Emotional	0.68	-0.42	0.68 (.57–.79)	15	0.74	0.53	0.54
Total	0.80	-0.59	0.82 (.72–.92)	40	0.80	0.79	0.79

Quantitative Evaluation of Bias

A series of studies using multigroup bifactor confirmatory factor analysis (MG-CFA) were conducted to evaluate potential bias at the item and test level. Items are statistically free from bias when the factor structure, item-factor loadings, and item intercepts are not statistically different across groups. This methodology is typically referred to as measurement equivalence/invariance. Here we summarize the results of those studies. See the following publications for a full description of each study.

- Across race/ethnicity in Pendergast, von der Embse, Kilgus, & Eklund, 2017
- By gender in von der Embse, Kilgus, Iaccarino, & Levi-Nielsen, 2017
- By race and gender using student level screening data from the FastBridge system

Approach

Item and test bias were examined using MG-CFA. Specifically, the overall fit for a series of nested MG-CFA models in which the models progress from least constrained to most constrained was used. The model progression is shown below.

- Evaluate invariance in the overall factor structure (three subdomain—one broad domain bifactor model) across groups.
- Evaluate invariance of factor loadings across groups—referred to as the “equal loadings” model.
- Evaluate invariance of intercepts across groups—referred to as the “equal intercepts” model.
- Evaluate invariance of residuals across groups—referred to as the “equal residuals” model.

Following well-established standards for evaluating bias using MG-CFA, minimally acceptable model fit was defined a priori as: RMSEA values ≤ 0.08 and CFA values ≥ 0.90 (Browne & Cudeck, 1993; Hu & Bentler, 1995; Markland, 2007). Measurement invariance of the two-factor SAEBRS structure was assessed by

applying increasingly restrictive equality constraints across groups to examine (a) configural invariance, (b) metric invariance, and (c) scalar/threshold invariance. Nested models (i.e., models with increasingly restrictive invariance tests) were compared using the change in Satorra-Bentler (SB) chi-square χ2 (ΔSB χ2), change in the comparative fit index (ΔCFI), and change in root mean square error of approximation (ΔRMSEA) values. Within each study, each nested model was compared to its parent model, the latter of which possessed increasingly restrictive invariance specifications. As the models grew more restrictive, the change in CFI (ΔCFI) (Cheung & Rensvold, 2002) and the change in RMSEA (ΔRMSEA) were used to evaluate comparability of the models. Small changes in fit indicate comparability across models (Byrne, 2011; Meade, et al., 2008; Satorra & Bentler, 2001).

Study 1

Study 1 evaluated bias on an early edition of SAEBRS, called SABRS, which did not include the emotional scale items. Data were collected in an elementary and middle school in the Southeastern United States. Teachers rated 864 students. Fifty-two percent of the sample was female, 48% White, 37% African American, and the remaining students representing other racial groups. The sample included students from grades 1–9 with a fairly even distribution between elementary and middle school grades.

Results are presented in Table 37. The first two models are based on separate analysis by race: African American, and White. The third model evaluates equivalence of the factor structure. The fourth model indicates equivalence of factor loadings. The fifth model indicates equivalence of the intercept/threshold.

Model fit was very good for all models, exceeding the thresholds of >0.90 CFI and <0.08 RMSEA. The changes in model fit with progressively restrictive models were very small and not significant, indicating that the slopes and intercepts are statistically equivalent across groups and no bias was present.

Table 37

Model Invariance on SABRS by Race

Model	CFI	ΔCFI	RMSEA	ΔRMSEA
African American	0.993	-	0.085	-
White	0.993	-	0.070	-
Factor Structure Equivalence	0.993	-	0.077	-
Item Loading Equivalence	0.994	0.001	0.069	0.008
Item Intercept Equivalence	0.994	0.0	0.062	0.007

Study 2

This study evaluated MG-CFA invariance on mySAEBRS using self-ratings from a sample of 1,102 students from two middle schools in a large urban district. In the sample 45% identified as female, 45% as male, and 10% other or not available. Thirty-one percent identified as White, 31% African American, and the remaining 48% from other racial groups.

Measurement invariance evaluated possible gender bias. A three-factor MG-CFA model was fit to the data. Results are reported in Table 38. Model fit indices indicate very strong fit to the data and the change indices indicate comparability between groups across models. Thus, no differential functioning or bias was observed.

Table 38

Model Invariance on mySAEBRS by Gender.

Model	CFI	ΔCFI	RMSEA	ΔRMSEA
Factor Structure Equivalence	0.998	-	0.058	-
Item Loading Equivalence	0.999	0.001	0.037	0.021
Item Intercept Equivalence	0.999	0.0	0.037	0.000

FastBridge Screening Data Study

Two additional studies were conducted to evaluate potential bias by gender and race on SAEBRS and mySAEBRS using deidentified FastBridge customer screening data from the fall of 2019. For these studies, a sample of 10,000 students was randomly selected such that the sample had an equal percentage of males and females and the composition by race group closely matched the U.S. school population. The sample consisted of 5,800 White students, 1,700 Hispanic, 1,600 African American, and 900 from other race groups. The sample was evenly distributed across grades 3–8 with about 1,500 per grade. About 1,000 grade 2 students were also included. Because students were randomly sampled from the entire database, the data come from schools across the U.S. For both the SAEBRS and mySAEBRS samples, data from more than 500 schools were used.

The MG-CFA invariance approach described above was used to evaluate possible bias by gender and by race. Data from three race groups, African American, Hispanic, and a combination of other races were evaluated in relation to Whites. Results for SAEBRS are reported in Table 38 and for mySAEBRS in Table 39.

The overall model of each model to the SAEBRS teacher ratings was good. RMSEA values ranged from 0.08 to 0.087 and CFI values were near 0.92. Comparisons across nested models resulted in very small changes in RMSEA and CFI—thus supporting the conclusion that SAEBRS ratings are invariant across gender and race groups.

Table 39

SAEBRS MG-CFA Model Invariance by Comparison Groups

Comparison	Model	RMSEA	CFI	ΔRMSEA	ΔCFI
Female Male	Factor Structure Equivalence	0.086	0.921	-	-
	Item Loadings Equivalence	0.083	0.918	0.003	0.003
	Item Intercept Equivalence	0.082	0.915	0.001	0.003
Comparison	Model	RMSEA	CFI	ΔRMSEA	d CFI
White Hispanic	Factor Structure Equivalence	0.087	0.920	-	-
	Item Loadings Equivalence	0.083	0.919	0.004	0.001
	Item Intercept Equivalence	0.081	0.918	0.002	0.001
Comparison	Model	RMSEA	CFI	ΔRMSEA	d CFI
White African Amer.	Factor Structure Equivalence	0.088	0.920	-	-
	Item Loadings Equivalence	0.084	0.918	0.004	0.002
	Item Intercept Equivalence	0.082	0.917	0.002	0.001

Comparison	Model	RMSEA	CFI	ΔRMSEA	d CFI
White African Amer.	Factor Structure Equivalence	0.087	0.921	-	-
	Item Loadings Equivalence	0.082	0.920	0.005	0.001
	Item Intercept Equivalence	0.080	0.920	0.002	0.000

The overall fit of the 3-factor model to mySAEBRS ratings is mixed. RMSEA values indicate good fit (<0.08); however, the CFI values indicate modest fit as they are below the threshold for good fit: 0.92. The marginal CFI results may be due to one of the academic items that had poor fit in the overall sample. However, given the strong fit indicated by the RMSEA, interpretation of the change in the fit is appropriate. In all comparisons, the small change across models supports the conclusion that the model and its parameters are invariant across groups and no statistical bias is present.

Table 40

mySAEBRS MG-CFA Model Invariance by Comparison Groups

Comparison	Model	RMSEA	CFI	ΔRMSEA	ΔCFI
Female Male	Factor Structure Equivalence	0.062	0.843	-	-
	Item Loadings Equivalence	0.059	0.842	0.003	0.001
	Item Intercept Equivalence	0.059	0.833	0.000	0.009
Comparison	Model	RMSEA	CFI	ΔRMSEA	d CFI
White Hispanic	Factor Structure Equivalence	0.062	0.850	-	-
	Item Loadings Equivalence	0.060	0.842	0.002	0.008
	Item Intercept Equivalence	0.060	0.838	0.000	0.004
Comparison	Model	RMSEA	CFI	ΔRMSEA	d CFI
White African Amer.	Factor Structure Equivalence	0.061	0.849		
	Item Loadings Equivalence	0.059	0.845	0.002	0.004
	Item Intercept Equivalence	0.058	0.841	0.001	0.004
Comparison	Model	RMSEA	CFI	ΔRMSEA	d CFI
White African Amer.	Factor Structure Equivalence	0.061	0.854		
	Item Loadings Equivalence	0.060	0.845	0.001	0.009
	Item Intercept Equivalence	0.057	0.851	0.003	-0.006

earlyReading

Overview

The FastBridge earlyReading assessment system consists of a set of 13 brief measures that assess specific foundational reading skills in the areas of concepts of print, phonemic awareness, phonics, decoding, and fluency in kindergarten and grade 1. For each grade and season, a subset of four are recommended for universal screening. The recommended set follows the common sequence of reading skills development in these grades and are combined into a composite score. The composite score is weighted to provide the strongest prediction of general reading proficiency and reading success by third grade. The individual measures provide important diagnostic information about each student's individual reading skill development and can be used to establish performance goals. For each skill, 20 alternate forms are available and can be used to reliably monitor student progress toward goals on a weekly basis. The table below lists the 13 earlyReading measures, the reading skill assessed, and whether it is included in the earlyReading composite score.

Table 41

earlyReading Assessments

Subtest	Reading Skill	earlyReading Composite
Concepts of Print*	Concepts of Print	Fall, Kindergarten
Letter Names*	Alphabetic Principle (Phonics)	Fall, Kindergarten
Letter Sounds*	Alphabetic Principle (Phonics)	Fall–Spring, Kindergarten
Decodable Words	Alphabetic Principle (Phonics)	--
Nonsense Words*	Alphabetic Principle (Phonics)	Winter, Kindergarten; Spring, Grade 1
Sight Words (50)*	Fluency	Spring, Kindergarten
Sight Words (150)*	Fluency	Fall–Spring, Grade 1
Sentence Reading*	Fluency	Fall, Grade 1
CBMreading	Fluency	Winter–Spring, Grade 1
Onset Sounds	Phonemic Awareness	Fall–Winter, Kindergarten
Rhyming	Phonemic Awareness	--
Word Blending	Phonemic Awareness	--
Word Segmenting*	Phonemic Awareness	Winter, Kindergarten–Spring, Grade 1
Oral Repetition	Phonemic Awareness	--

Reliability

There are several forms of reliability evidence, including internal-consistency, alternate-form, and inter-rater, that address different sources or random error that impact score consistency across time. The choice of evidence depends in part on the purpose of the assessment and the nature of the assessment. Internal-consistency reliability is generally appropriate for assessments comprised of multiple tasks or items that are combined into a single score. The most common form in education is Cronbach's alpha, but it is only appropriate for non-speeded assessments that are unidimensional and that assess a single skill.

Several earlyReading assessments use relatively short time limits to generate rate-based scores that represent both accuracy and automaticity; therefore, Cronbach's alpha is not appropriate. Because these assessments are also used for frequent progress monitoring and include up to 20 unique forms, alternate-form reliability is a better choice. Alternate-form reliability coefficients tend to be lower than internal-consistency coefficients because they are impacted by additional sources of random error, including error due to minor differences in content coverage, and time.

Cronbach's alpha is not appropriate for estimating the reliability of the earlyReading composite score because each composite consists of subtests that assess different skills. A better alternative is a confirmatory factor model-based approach. With this approach, a model is constructed in which each subtest loads onto a single factor. This model-based approach is a form of internal-consistency reliability.

Most of the earlyReading measures use open-ended performance tasks, which require some level of judgement by the examiner to determine the correctness of the response. For these assessments, inter-rater reliability is especially relevant.

Finally, delayed-retest reliability was computed for the earlyReading composite. This form of reliability is useful when a student needs to be reassessed on the same content within a season, as can happen when a student changes schools, or when the first administration was spoiled.

Internal-Consistency Reliability

Internal-consistency Cronbach's alpha reliability of the untimed earlyReading assessments were calculated on a random sample of students from the 2012–2013 academic year.

Table 42

Cronbach's Alpha for Untimed earlyReading Assessments

Subtest	Grade	N	Coefficient	CI
Concepts of Print	K	336	0.75	0.70–0.79
Onset Sounds	K	597	0.87	0.85–0.89
Rhyming	K	586	0.94	0.93–0.95
Word Blending	K–1	480	0.90	0.88–0.92
Word Segmenting	K–1	500	0.95	0.94–0.96

Internal-consistency reliability of the earlyReading composite score was evaluated using a confirmatory factor analysis model. For each grade and season, a single factor model was defined by regressing the latent composite score onto the four subtests comprising the composite. Reliability is computed as the ratio of the sum of the squared factor loadings (true score variance) divided by the total variance.

$$\frac{\sum_i^k \lambda_i^2}{\sum_i^k \lambda_i^2 + \sigma_e^s}$$

where lambda (λ) is the factor loading and error variance is the sum of 1-λ^2

earlyReading universal screening scores of more than 700 schools during the 2017–2018 academic school year were used. The demographic composition of the sample is shown in Table 43. Internal-consistency reliability is strong with all coefficients in the 0.80s.

Table 43

Internal Consistency Reliability of the earlyReading Composite

					Percent of Sample				
Grade	Season	N	Rxx	95% CI	Female	White	A.A.	Hisp.	Other
K	Fall	47,475	0.86	0.86–0.86	49	71	6	12	10
K	Winter	42,036	0.83	0.83–0.83	49	71	6	12	10
K	Spring	42,378	0.83	0.83–0.83	49	71	6	12	10
1	Fall	38,885	0.87	0.87–0.87	48	73	6	10	11
1	Winter	39,073	0.84	0.84–0.84	48	73	6	10	11
1	Spring	39,208	0.85	0.85–0.85	48	73	6	10	11

Alternate-Form Reliability

Alternate-form reliability was used to estimate reliability for the speeded earlyReading measures. Using progress monitoring scores collected during the 2018–2019 school year, the correlation between pairs of forms administered in sequence was computed. The median correlation coefficient across the 20 correlated pairs is used as an overall estimate of alternate-form reliability. The sample consists of students from 15 states in each grade with representation from all race/ethnicity categories; urban, suburban, and rural districts; and the full range of free/reduced lunch rates. Median coefficients range from 0.70 to 0.91, with most in the upper 0.70s to mid 0.80s. Coefficients in this range indicate strong score stability across administrations.

Table 44

Alternate-Form Reliability for earlyReading Assessments.

Measure	Grade	N	Median	CI
Letter Names	K	7,993	0.85	0.84–0.86
Letter Sounds	K	30,725	0.82	0.82–082
Sight Words (50)	K	2,013	0.91	0.90–0.91
Sight Words (150)	1	8,833	0.88	0.88–0.89
Decodable Words	K	2,335	0.76	0.75–0.78
Decodable Words	1	5,614	0.70	0.69–0.71
Nonsense Words	K	5,905	0.81	0.80–0.82
Nonsense Words	1	17,072	0.77	0.77–0.78

Inter-rater

Data for the inter-rater reliability analysis came from administrations done in 2014 by experienced graduate students who attended a two-hour training session. Assessments were administered to

students in nine elementary schools within three school districts in a metropolitan area in the Midwest. Students were administered five randomly selected progress monitoring forms. District A was about 56% White, 14% African American, 10% Hispanic, and 19% Asian/Pacific Islander. About 45% of students were eligible to receive free/reduced lunch and 13% were eligible for special education services. District B was about 93% White, 4% African American, 3% Hispanic, and 4% Asian/Pacific Islander. About 17% of students were eligible to receive free/reduced lunch and 10% were eligible for special education services. District C was about 80% White, 7% African American, 5% Hispanic, and 11% Asian/Pacific Islander. About 45% of students were eligible to receive free/reduced lunch and 10% were eligible for special education services.

Table 45

Inter-Rater Reliability of earlyReading Assessments

Measure	Grade	Correlation	N
Onset Sounds	K	0.98	40
Letter Sounds	K	0.99	47
Letter Names	K	0.99	69
Word Blending	K	0.98	95
Word Blending	1	0.89	159
Word Segmenting	K	0.85	90
Word Segmenting	1	0.83	85
Nonsense Words	1	0.99	51
Decodable Words	1	0.99	120
Sight Words (50)	K	0.99	9
Sight Words (150)	1	0.97	125

Test-retest Reliability

Test-retest reliability evidence was derived from the rescreening of students a second time within a season during the 2017–2018 school year. Only students who were retested within 60 days were included in the analysis. Table 46 reports results for each measure used in the composite and for the earlyReading composite.

Table 46

Test-Retest Reliability of earlyReading Measures

Measure	Grade	Season	N	Rxx
Concepts of Print	K	Fall	342	0.72
Onset Sounds	K	Winter	317	0.78
Letter Names	K	Fall	471	0.81
Letter Sounds	K	Spring	240	0.85
Word Segmenting	K	Spring	193	0.78
Sight Words (50)	K	Spring	177	0.84
Nonsense Words	1	Fall	197	0.85
Sight Words (150)	1	Fall	220	0.88
Sentence Reading	1	Fall	211	0.94

CBMreading	1	Spring	268	0.97
Composite	K	Fall	312	0.86
Composite	K	Winter	257	0.85
Composite	K	Spring	142	0.90
Composite	1	Fall	204	0.95

Validity

Content Validity

The FastBridge earlyReading system assesses key foundational reading skills in the areas of concepts of print, phonics, phonemic awareness, and fluency. These areas are cited in the National Reading Panel Report (NRP, 2000) and they appear in all 50 states' early reading learning standards. Additionally, the evidence of content validity draws from a large research base that demonstrates the critical roles these skills play in typical reading development. See Table 41 for a list of reading skills assessed by each earlyReading measure.

Criterion Validity

Criterion-related validity of earlyReading measures and composites was examined using correlational evidence with key indicators of general reading ability. In the first study, scores on earlyReading measures were correlated with scores on the Group Reading Assessment Classification Evaluation (GRADE). The GRADE is an untimed, group-administered, norm-referenced reading achievement test that is intended for children in preschool through grade 12. Comprised of 16 subtests categorized within five components, the GRADE utilizes subtest scores, depending on the testing level, to form the Total Test composite score. Evidence for the validity of earlyReading is presented below on the external criterion measure of the GRADE Total Test composite score.

To establish criterion-related validity, students were recruited from school districts. In School District 1, three elementary schools participated. Kindergarten students from District 1 who participated in the study were enrolled in all-day or half-day kindergarten. Most students within the school district were White (78%), with the remaining students identified as either African American (19%) or other (3%). Forty to fifty percent of students at each school were eligible for free/reduced lunch. In School District 2, most students within the school district were White (53%), with the remaining students identified as African American (26%), Hispanic (11%), Asian (8%), or other (2%). Forty to fifty percent of students at each school are on free/reduced lunch.

A summary of concurrent and predictive validity coefficients for the GRADE is presented in Table 47. Scores from the fall, winter, and spring administrations of earlyReading were correlated with a spring administration of GRADE and are represented by the coefficient range and median. As discussed previously, the composite score in kindergarten and first grade demonstrated the highest level of criterion validity, which suggests that it is the best estimate of current and later broad reading performance.

Table 47

Concurrent and Predictive Validity of earlyReading Measures with GRADE Total Score

Subtest	Grade	N (range)	Coefficient (range)	Coefficient Median
Composite	K	173	0.67–.69	0.68
Composite	1	100	0.72–.83	0.81
Onset Sounds	K	85–230	0.03–.62	0.58
Letter Names	K	85–230	0.18–.63	0.44
Letter Sounds	K	85–230	0.19–.63	0.49
Word Blending	K	213–230	0.23–.66	0.41
Word Blending	1	71–179	0.12–.56	0.38
Word Segmenting	K	213–228	0.25–.58	0.42
Word Segmenting	1	71–179	0.07–.60	0.41
Decodable Words	K	214	-	0.27
Decodable Words	1	71–179	0.22–.78	0.53
Sight Words-50	K	213	-	0.19
Sight Words-150	1	71–179	0.43–.80	0.66
Nonsense Words	K	105–215	0.27–.44	0.36
Nonsense Words	1	168–179	0.43–.67	0.60

Additional validity evidence for select earlyReading measures and the earlyReading composite was obtained from universal screening data with large nationally representative student populations. For each earlyReading measure, another earlyReading measure that represented a later and more general stage in early reading skill development was used as the criterion. This is ideal because reading development builds upon prior skills and readers must achieve mastery or automaticity on the more basic skills to be successful. As an example, Onset Sounds is a very basic form of phonemic awareness in which a student must recognize the first sound in a word. The criterion measure, Word Segmenting, constitutes a broader and more complex form of phonemic awareness in which the reader must be able to identify all phonemes in a word.

Concurrent validity evidence used data from a single screening period. Results are presented in Table 48. Letter Sounds was used as the criterion for letter names, Word Segmenting was used as the criterion for Onset Sounds and Letter Sounds, and CBMreading words correct per minute were used as the criterion for Nonsense Words and Sight Words. The FastBridge adaptive reading assessment serves as the criterion measure for the earlyReading composite score. aReading is a standards-based assessment that measures state literacy standards. Unlike the earlyReading measures that use rate-based scores to assess automaticity, aReading is untimed and assesses the full range of early literacy and reading skills, from foundational reading standards through reading comprehension.

Additionally, aReading has been validated as a strong predictor of end-of-grade state assessments (see aReading validity sections above).

Validity coefficients of the earlyReading measures are expected to be moderate, in the 0.50s to 0.60s, due to narrow skills focus and the brevity of the measures. Concurrent validity results reported in Table 48 are consistent with expectations or better than expectations. The validity coefficients for Onset Sounds and Letter Sounds are moderate, whereas the coefficients for Nonsense Words and Sight Words are strong. Onset Sounds and Letter Sounds measure a narrow aspect of the sound phoneme correspondence and are usually fully developed by the time students can fully segment words. So, although they are critical foundational reading skills, they are not strong predictors of broad word reading ability. The strength of the association of Nonsense Words and Sight Words with CBMreading is expected because all three involve the ability to decode words. Moreover, the FastBridge CBMreading measure in grade 1 exclusively uses decodable and sight words.

Concurrent validity coefficients for the earlyReading composite with aReading are strong and range from 0.66 in the fall of kindergarten to 0.78 in the winter and spring of grade 1 (Table 49). The National Center on Intensive Intervention (NCII) considers validity coefficients of 0.60 and higher as providing strong evidence (National Center on Intensive Intervention, 2020).

Table 48

Concurrent Validity of Select earlyReading Measures

Measure	Grade	N	Coefficient	CI Lower Bound	CI Upper Bound	Criterion
Onset Sounds	K	42,036	0.57	0.57	0.57	Word Segmenting
Letter Names	K	47,475	0.77	0.77	0.77	Letter Sounds
Letter Sounds	K	42,036	0.51	0.51	0.51	Word Segmenting
Nonsense Words	1	39,208	0.77	0.77	0.77	CBMreading
Sight Words	1	39,208	0.84	0.84	0.84	CBMreading

Table 49

Concurrent Validity of the earlyReading Composite with aReading

Grade	Season	N	Coefficient	CI Lower Bound	CI Upper Bound	Criterion
K	Fall	9,779	0.66	0.65	0.67	aReading
K	Winter	32,303	0.67	0.67	0.67	aReading
K	Spring	33,135	0.72	0.72	0.72	aReading
1	Fall	32,214	0.74	0.74	0.74	aReading
1	Winter	32,303	0.78	0.78	0.78	aReading
1	Spring	33,135	0.78	0.78	0.78	aReading

Predictive validity coefficients are based on data collected in two screening periods. For example, the coefficient for Onset Sounds is based on winter Onset Sounds scores correlated with Spring Word Segmenting scores. Results for the individual earlyReading measures are reported in Table 50, which also indicates, in parentheses, the season the score was collected. Table 51 shows results for the earlyReading composite score.

Predictive validity coefficients are generally strong ranging from 0.61 for Letter Names to 0.83 for Sight Words. For Onset Sounds and Letter Sounds, the predictive validity coefficient is stronger than the concurrent validity coefficient. This difference may demonstrate the effects of the lag in the development of word segmenting skills relative to onset sounds and letter sounds skills. Predictive validity coefficients for the earlyReading composite were also strong.

Table 50

Predictive Validity of Select earlyReading Measures

Measure	Grade	N	Coefficient	CI Lower Bound	CI Upper Bound	Criterion
Onset Sounds (W)	K	41,536	0.66	0.66	0.66	Word Segmenting (S)
Letter Names (F)	K	40,508	0.61	0.61	0.61	Letter Sounds (W)
Letter Sounds (W)	K	41,536	0.70	0.70	0.70	Word Segmenting (S)
Nonsense Words (F)	1	37,533	0.72	0.72	0.72	CBMreading (S)
Sight Words (F)	1	37,533	0.83	0.83	0.83	CBMreading (S)

Table 51

Predictive Validity of the earlyReading Composite with aReading, Spring

Grade	Season	N	Coefficient	CI Lower Bound	CI Upper Bound	Criterion
K	Fall	8,128	0.66	0.65	0.67	aReading
K	Winter	20,238	0.70	0.70	0.70	aReading
1	Fall	26,690	0.70	0.70	0.70	aReading
1	Winter	28,958	0.77	0.77	0.77	aReading

Classification Accuracy

Two studies evaluating classification accuracy of the earlyReading composite score in kindergarten and grade 1 were conducted: one using the GRADE as the criterion, the other using aReading as the criterion. The samples are the same as those described in the validity sections above. For both studies,

the 40th national percentile in the spring was used as the criterion level divided at-risk students below from not at-risk studies above the threshold.

Several indices are used to evaluate the quality of classification accuracy. AUC values above 0.90 indicate a strong classifier. In the study with GRADE as the criterion measure, all AUCs exceed 0.90. In the much larger study with aReading as the criterion, the evidence from AUC indicates a moderately strong classifier. A generally accepted criterion for evaluating the classification accuracy of a brief screener is sensitivity and specificity values of 0.70 or higher. Sensitivity exceeds the threshold in all but two instances; whereas, and specificity exceed the threshold in all instances. Other things being equal, in education higher sensitivities even at the expense of specificity are preferred because greater sensitivity means that fewer students slip through the cracks.

Table 52

Classification Accuracy of the earlyReading Composite with GRADE & aReading

Grade	Criterion	AUC	Sensitivity	Specificity	Overall Accuracy
K					
Fall	GRADE	0.91	0.88	0.88	0.84
Winter	GRADE	0.91	0.94	0.72	0.72
Spring	GRADE	0.95	0.75	0.74	0.74
Grade 1					
Fall	GRADE	0.98	1.00	0.93	0.93
Winter	GRADE	0.98	1.00	0.82	0.83
Spring	GRADE	0.99	0.89	0.90	0.90
K					
Fall	aReading	0.78	0.85	0.59	0.78
Winter	aReading	0.87	0.83	0.63	0.87
Spring	aReading	1.00	0.79	0.71	0.76
Grade 1					
Fall	aReading	0.87	0.73	0.87	0.87
Winter	aReading	0.93	0.82	0.84	0.93
Spring	aReading	1.00	0.87	0.79	0.83

earlyMath

Overview

FastBridge earlyMath is a suite of 17 brief assessments for universal screening, diagnostic evaluation of strengths and weaknesses, and monitoring student progress on early numeracy skills. For universal screening, the earlyMath assessments are recommended to be used with students in kindergarten and grade 1. The results from three earlyMath assessments are combined into a composite score that serves as the primary data point for the overall evaluation of risk. In this way, the earlyMath composite score provides an efficient, instructionally relevant, and technically adequate assessment of early math skill development and serves as an early indicator of potential math difficulties. For diagnostic evaluation and progress monitoring, these assessments can be used with students from any grade who are developing these skills. For frequent progress monitoring, 20 alternate forms are available per skill. Table 53 provides a complete list of the earlyMath measures, the associated numeracy domain, and whether the measure is included in the earlyMath composite. Those with NA are optional measures that teachers can use to obtain further diagnostic information about a student's learning needs.

Table 53

earlyMath Measures

Measure	Early Numeracy Domain	earlyMath Composite
Match Quantity	Counting & Cardinality	Fall, Kindergarten
Numeral Identification (K)	Counting & Cardinality	Fall–Spring, Kindergarten
Number Sequence (K)	Counting & Cardinality	Fall–Spring, Kindergarten
Decomposing (K)	Number & Operations: Base 10	Winter–Spring, Kindergarten
Subitizing	Counting & Cardinality	--
Quantity Discrimination (Most/Least)	Counting & Cardinality	--
Composing	Number & Operations: Base 10	--
Counting Objects	Counting & Cardinality	--
Equal Partitioning	Counting & Cardinality	--
Numeral Identification (1)	Number & Operations: Base 10	Fall, Grade 1
Decomposing (1)	Operations & Algebraic Thinking	Fall–Spring, Grade 1
Number Sequence (1)	Number & Operations: Base 10	Winter–Spring, Grade 1
Place Value	Number & Operations: Base 10	Winter–Spring, Grade 1
Story Problems	Number & Operations: Base 10	Spring, Grade 1
Verbal Addition	Operations & Algebraic Thinking	--
Verbal Subtraction	Operations & Algebraic Thinking	--

Reliability

As with the earlyReading CBM, different forms of reliability are pertinent to the various earlyMath measures. For untimed, non-speeded measures, Cronbach's alpha is used to estimate internal-consistency reliability. For the speeded measures of automaticity, alternate-form reliability is used. A confirmatory factor model-based approach is used to evaluate reliability of the earyMath composite scale. Inter-rater and retest reliability are also evaluated on select measures.

Internal-Consistency Reliability

Internal consistency of Cronbach's alpha reliability of the untimed earlyMath assessments was calculated on a random sample of students from the 2018–2019 academic year.

Table 54

Cronbach's Alpha for Untimed earlyMath Measures

Measure	Grade	N	Coefficient	Confidence Interval
Composing	K	603	0.82	0.79–0.84
Subitizing	K	854	0.83	0.81–0.85

Counting Objects	K	76	0.83	0.74–0.89
Decomposing	K	601	0.80	0.77–0.83
Equal Partitioning	K	236	0.71	0.64–0.77
Number Sequence	K	598	0.76	0.72–0.79
Number Sequence	1	572	0.85	0.83–0.87
Story Problems	1	138	0.81	0.74–0.86
Place Value	1	91	0.82	0.74–0.88
Decomposing	1	573	0.87	0.85–0.89

Alternate-Form Reliability

Alternate-form reliability was used to estimate reliability for earlyMath assessments that include alternate progress monitoring forms and for the speeded measures for which internal consistency is not appropriate. Using progress monitoring scores collected during the 2018–2019 school year, the correlation between pairs of forms administered in sequence was computed. The median correlation coefficient is based on a minimum of four forms collected within a 30-day interval (approximately 7 days between forms) and is used as an overall estimate of alternate for reliability. The sample consists of students from 15 states in each grade with representation from all race/ethnicity categories; urban, suburban, and rural districts; and the full range of free/reduced lunch rates. Median coefficients range from 0.76 to 0.91, with most in the mid 0.80s. Coefficients in this range indicate strong score stability across administrations.

Table 55

Alternate-Form Reliability for earlyMath Assessments

Measure	Grade	N	Median	CI
Match Quantity	K	1,200	0.83	0.81–0.85
Number Identification	K	8,535	0.91	0.91–0.91
Number Sequence	K	3,770	0.84	0.83–0.85
Quantity Discrimination (Most)	K–1	140	0.85	0.80–0.89

Place Value	1	708	0.86	0.84–0.88
Decomposing	1	5,790	0.79	0.78–0.80
Number Identification	1	2,027	0.90	0.89–0.91
Number Sequence	1	828	0.76	0.73–0.79

Test-Retest Reliability

Test-retest correlations between student scores on two different testing sessions were calculated in January 2014. Experienced graduate students who attended a two-hour training session collected the data. After the first testing session, the same data collectors then administered the earlyMath measures to the same individual students one to three weeks later. We collected Place Value subtest data using group administration procedures. Examiners tested students by classrooms to be more efficient and to limit distractions to the teachers' classrooms.

Reliability coefficients across subtests were moderate to high, ranging from $r = 0.62$ to 0.87 in kindergarten and $r = 0.71$ to 0.91 in first grade. In both grades, the test-retest reliability coefficients for the earlyMath composite score were highest, indicating that it may be the most stable measure of a student's early numeracy skills. The sample consisted of 39 students in kindergarten and 39 students in first grade. Approximately 65% of students in kindergarten and 49% of students in first grade were female. The ethnicity breakdown across both grades was 59%–82% White, 10%–21% African American, 5%–8% Hispanic, 0%–13% Asian/Pacific Islander, and 0%–3% American Indian or Alaskan Native. In addition, 3%–5% of students were receiving special education services and 15%–38% of students were eligible for free/reduced lunch.

Table 56

Test-retest Reliability of earlyReading measures

Measure	Grade	N	Rxx
Match Quantity	K	37	0.87
Numeral Identification	K	36	0.76
Number Sequence	K	38	0.85
Decomposing	K	35	0.80
Counting Objects	K	38	0.68
Equal Partitioning	K	33	0.68
Composing	K	32	0.71
Quantity Discrimination (Most)	K	38	0.62
Quantity Discrimination (Least)	K	38	0.73
Numeral Identification	1	36	0.91
Number Sequence	1	36	0.71

Decomposing	1	36	0.83
Place Value	1	57	0.77
Story Problems	1	39	0.77
Composite	**K**	**37**	**0.87**
Composite	**1**	**30**	**0.91**

Inter-rater Reliability

Inter-rater reliability is a measure of the extent to which student scores are consistent across different examiners or scorers. Estimates of inter-rater reliability are based on two independent scorers, and the coefficients represent the level of agreement between examiners or raters. earlyMath measures involve a small degree of subjectivity, given the clear scoring guidelines. However, unreliable scoring may be the result of timing discrepancies between administrators or scorers or differences in the interpretation of a student's response. A summary of inter-rater reliability is presented in Table 57. When low agreement was observed across raters it was typically the result of hesitation or discontinue rules. Overall, the median inter-rater reliability across kindergarten was high (93% to 100% agreement).

Table 57

Inter-Rater Reliability of earlyMath Measures

Measure	Average Correlation	Median Correlation	Correlation Range	N
Kindergarten				
Match Quantity	0.93	0.93	0.58–1.00	45
Numeral Identification	0.96	0.98	0.83–1.00	45
Number Sequence	0.99	1.00	0.85–1.00	45
Decomposing	0.95	1.00	0.75–1.00	45
Composing	0.95	1.00	0.63–1.00	45
Quantity Discrimination (Most)	0.91	0.93	0.47–1.00	45
Quantity Discrimination (Least)	0.91	0.93	0.47–1.00	45
First Grade				
Numeral Identification	0.98	0.98	0.92–1.00	45

Number Sequence	0.97	1.00	0.71–1.00	45
Decomposing	0.93	1.00	0.67–1.00	45
Place Value	0.98	1.00	0.75–1.00	62
Story Problems	0.94	1.00	0.67–1.00	45

Validity

Content Validity

The FastBridge earlyMath system assesses early numeracy and other key foundational math concepts and operation skills. The areas assessed align with state and national learning standards and reflect the developmental progression from counting and cardinality to base-10 concepts and basic addition and subtraction skills and operations concepts. A complete description of the content assessed in the earlyMath system is provided in the *FastBridge Assessments Content Description and Use Guide V5* (Illuminate Education, 2020).

Criterion Validity

Criterion-related validity of earlyMath measures and the composite score was examined using correlational evidence with key indicators of general math achievement: (a) Measures of Academic Progress for Primary Grades (MAP PG), and (b) Group Mathematics Assessment and Classification Evaluation (GMADE; Level R and 1). The MAP (Northwest Evaluation Association, 2005) is a classification and computer-adaptive assessment designed to measure mathematics ability and progress. In kindergarten and first grade, MAP measures mathematics achievement in problem-solving, algebra, computation, measurement, statistics, and number sense. MAP data were collected in the winter and spring for kindergarten and in the fall, winter, and spring for first grade students at two schools located within one district in a suburban town as part of the district's regular screening process. The sample consisted of 221 students in kindergarten and 195 students in first grade. Approximately 50% of students were female. The ethnicity breakdown across both grades was 84%–91% White, 3%–4% African American, 2%–4% Hispanic, 2%–7% Asian/Pacific Islander, and 1%–3% American Indian or Alaskan Native. In addition, 7%–12% of students were receiving special education services and 30%–35% of students were eligible for free/reduced lunch.

The GMADE is a norm-referenced assessment of mathematics skills. Kindergarten students were administered Level R forms, which are composed of two subtest measures: Concepts and Communication, and Process and Applications. First grade students were administered Level 1 forms, which are composed of three subtest measures: Concepts and Communication, Process and Applications, and Operations and Computation. The assessment was administered to students in three school districts located in the Midwest near a metropolitan city. The GMADE was administered to students in two testing sessions by trained graduate students. The assessment was administered two to four weeks after the spring earlyMath assessment was administered. The sample consisted of 155 students in kindergarten and 170 students in first grade. Approximately 44% of students were female. The ethnicity breakdown across both grades was 74%–80% White, 8%–14% African American, 6% Hispanic, 4%–5% Asian/Pacific Islander, and 1%–3% American Indian or Alaskan Native. In addition, 8%–13% of students were receiving special education services and 29%–34% of students were eligible for free/reduced lunch.

A summary of concurrent and predictive validity coefficients for the MAP and GMADE are presented for both kindergarten and first grade in the tables below. As discussed previously, the composite measure in

kindergarten and first grade demonstrated the highest level of criterion validity—suggesting that it is the best estimate of current and later broad mathematics performance.

Table 58

Concurrent & Predictive Validity Coefficients of earlyMath Measures with GMADE & MAP, Kindergarten

earlyMath Measure	Criterion	Corr. Range	Corr. Median
Composite	MAP	0.59–0.72	0.69
	GMADE	0.56–0.63	0.56
Match Quantity	MAP	0.39–0.57	0.52
	GMADE	0.34–0.47	0.44
Numeral Identification	MAP	0.43–0.64	0.58
	GMADE	0.46–0.51	0.47
Number Sequence	MAP	0.53–0.70	0.61
	GMADE	0.48–0.54	0.49
Decomposing	MAP	0.46–0.52	0.47
	GMADE	0.43–0.44	0.44
Counting Objects	MAP	0.36–0.37	0.37
	GMADE	-	0.25
Equal Partitioning	MAP	0.40–0.45	0.43
	GMADE	-	0.39
Composing	MAP	0.38–0.46	0.42
	GMADE	-	0.46
Quantity Discrimination (Most)	MAP	-	0.30
	GMADE	-	0.40
Quantity Discrimination (Least)	MAP	-	0.37
	GMADE	-	0.44

Table 59

Concurrent & Predictive Validity Coefficients of earlyMath Measures with GMADE & MAP, Grade 1

earlyMath Measure	Criterion	Corr. Range	Corr. Median
Composite	MAP	0.59–0.69	0.66
	GMADE	0.67–0.69	0.68
Numeral Identification	MAP	0.42–0.57	0.50
	GMADE	0.45–0.61	0.59
Number Sequence	MAP	0.58–0.64	0.64
	GMADE	0.56–0.65	0.60
Decomposing	MAP	0.51–0.59	0.56
	GMADE	0.56–0.63	0.59
Place Value	MAP	0.35–0.63	0.51
	GMADE	0.55–0.58	0.57
Story Problems	MAP	0.34–0.49	0.45
	GMADE	0.48–0.52	0.52

Additional criterion validity evidence was gathered for select earlyMath measures using large and diverse national samples. In these studies, other FastBridge assessments serve as criterion measures. Each criterion measure was selected because it measures skills that build upon the skills assessed in the focal measure or because it measured the broad range of concepts and skills the student should have acquired at the end of the grade. As an example of a criterion measure that builds on skills from the focal measure, Number Sequence was used as the criterion for Number Identification. Number Identification measures basic numeral recognition; Number Sequence requires the student to evaluate a sequence of numbers presented orally and determine the next or prior number in the sequence. The data were obtained as part of standard universal screening protocol in schools across the U.S. that includes urban, rural, and suburban districts and the full spectrum of race/ethnicity groups.

Table 60 provides concurrent validity evidence based on administrations conducted within the same season. The coefficients for the individual skill measures with other skills that assess similar or developmentally more complex math skills are moderate, with most coefficients in the 0.60s. That level of correlation is expected due to the narrowness and uniqueness of each skill and the brevity of the measures.

Table 61 provides concurrent validity evidence for the earlyMath composite with aMath by grade and season. As described above, the earlyMath composite combines scores from three individual earlyMath measures and serves as a broader indicator of overall math ability. The correlation coefficients with aMath are in the upper 0.60s and lower 0.70s for all but the fall of kindergarten. For a brief screener, validity coefficients of 0.60 are considered strong evidence (National Center on Intensive Intervention, 2020).

Table 60

Concurrent Validity of Select earlyMath Measures.

Measure	Grade	N	Coefficient	CI Lower Bound	CI Upper Bound	Criterion
Number Identification	K	268,630	0.69	0.69	0.69	Number Sequence (KG)
Number Sequence	K	137,424	0.61	0.61	0.62	Decomposing (KG)
Number Identification	1	42,036	0.71	0.71	0.72	Number Sequence (1)
Decomposing	1	14,509	0.63	0.62	0.64	aMath
Number Sequence	1	118,254	0.67	0.67	0.67	Number Sequence (1)
Place Value	1	10,740	0.52	0.52	0.52	aMath

Table 61

Concurrent Validity of the earlyMath Composite with aMath

Grade	Season	N	Coefficient	CI Lower Bound	CI Upper Bound	Criterion
K	Fall	3,914	0.62	0.60	0.64	aMath
K	Winter	7,155	0.67	0.66	0.68	aMath
K	Spring	8,727	0.68	0.67	0.69	aMath
1	Fall	9,285	0.72	0.71	0.73	aMath
1	Winter	10,833	0.74	0.73	0.75	aMath
1	Spring	12,963	0.71	0.70	0.72	aMath

Predictive validity evidence is provided in Tables 62 and 63. Validity coefficients are moderate and generally a bit lower than concurrent validity and consistent with expectations for brief individual skill measures. As a screener, the brief earlyMath composite provides strong prediction of end-of-grade achievement using a broad indicator of overall math ability.

Table 62

Predictive Validity of Select earlyMath Measures

Measure	Grade	N	Coefficient	CI Lower Bound	CI Upper Bound	Criterion
Number Identification	K	268,630	0.69	0.69	0.69	Number Sequence (K)
Number Sequence	K	137,424	0.61	0.61	0.62	Decomposing (K)
Number Identification	1	42,036	0.71	0.71	0.72	Number Sequence (1)
Decomposing	1	14,509	0.63	0.62	0.64	aMath
Number Sequence	1	118,254	0.67	0.67	0.67	Number Sequence (1)
Place Value	1	9,841	0.49	0.48	0.50	aMath

Table 63

Predictive Validity of the earlyMath Composite with aMath, Spring

Grade	Season	N	Coefficient	CI Lower Bound	CI Upper Bound	Criterion
K	Fall	3,914	0.60	0.58	0.62	aMath
K	Winter	7,155	0.67	0.66	0.68	aMath
1	Fall	9,285	0.71	0.70	0.72	aMath
1	Winter	10,833	0.76	0.75	0.77	aMath

Classification Accuracy

Three studies evaluating classification accuracy of the earlyMath composite score in kindergarten and grade 1 were conducted: one using the GMADE as the criterion, one using MAP, and one using aMath as the criterion. The samples are the same as those described in the validity sections above. For all three studies, the 40th national percentile in the spring was used as the criterion level divided at-risk students below from not-at-risk studies above the threshold. Criterion scores collected in the spring were used in each analysis. The results for GMADE across seasons were averaged to produce a single classification accuracy result.

Several indices were used to evaluate the quality of classification accuracy. AUC values above 0.90 indicate a strong classifier. The AUC values are generally in the 0.80s and provide good classification accuracy of the earlyMath composite with broad standards-based math assessments. A generally accepted criterion for evaluating the classification accuracy of a brief screener is sensitivity and specificity values of 0.70 or higher. In nearly all comparisons, sensitivity and specificity exceed this threshold. The only exception is the fall of kindergarten with aMath as the criterion.

Table 64

Classification Accuracy of the earlyMath Composite with GMADE, MAP, and aMath.

Grade	Criterion	N	AUC	Sensitivity	Specificity	Overall Accuracy
K	GMADE	121	0.86	0.81	0.82	0.83
Grade 1	GMADE	122	0.85	0.79	0.75	0.76
K						
Fall	MAP	259	0.80	0.74	0.72	0.74
Winter	MAP	266	0.87	0.76	0.76	0.84
Spring	MAP	247	0.88	0.81	0.88	0.76
Grade 1						
Fall	MAP	368	0.83	0.75	0.73	0.72
Winter	MAP	314	0.89	0.75	0.85	0.76
Spring	MAP	324	0.86	0.74	0.76	0.87
K						
Fall	aMath	3,914	0.76	0.69	0.68	0.69
Winter	aMath	7,155	0.85	0.71	0.78	0.75
Spring	aMath	8,727	1.00	0.71	0.80	0.75
Grade 1						

Fall	aMath	9,285	0.85	0.85	0.72	0.76
Winter	aMath	10,833	0.92	0.82	0.80	0.81
Spring	aMath	12,963	1.00	0.74	0.78	0.77

AUTOreading

Overview

FastBridge AUTOreading is a timed computer-based assessment consisting of eight subtests that measure the accuracy and rate of phonological (sound), orthographic (spelling), and semantic (meaning) skills. It was developed in response to requests from educators for fully automated assessments of foundational reading skills that can be used and interpreted across all grades, K–12. Subtest scores provide diagnostic information for students with poor reading comprehension skills. The subtest scores are also combined into a total AUTOreading composite score that is used alongside the aReading measure of general reading achievement in the FastBridge *Screening-to-Intervention* report. This report gives educators information for prioritizing reading instruction time and to determine instructional needs for whole classrooms, small groups, and individual students. AUTOreading also includes 20 alternate forms per test level and subtest level for frequent progress monitoring. Table 65 breaks out the skills by subtest, domain, and grade range.

Table 65

AUTOreading Skill and Domain Structure by Grade

Subtest	Domain	Grade Range		
		K	1–4	5–12
Letter Names	Phonics	✓		
Letter Sounds	Phonics	✓		
Encoding	Spelling		✓	
Word ID	Fluency		✓	✓
Vocabulary	Vocabulary		✓	
Synonym	Vocabulary			✓
Decoding	Phonics			✓
Morphology	Syntax/Comp.			✓

AUTOreading is scored using the IRT-based graded response model with four score levels 0–3 per item.

- 0 for incorrect
- 1 for correct and slow response time (bottom quartile)
- 2 for correct and moderate response time (25th–70th percentile)
- 3 for correct and fast response time (71st–99th percentile)

The resulting IRT-theta score is converted to the aReading scale. Table 66 provides summary statistics of the AUTOreading composite score on the aReading scale by grade. About 70% of the sample was tested in the fall with the other 30% tested during winter screening. As expected, the mean of the rescaled

AUTOreading scores increases with grade. Because these results were based on a convenience sample, the results do not generalize to the population of students.

Table 66

Rescaled AUTOreading Summary Statistics by Grade

Grade	N	Mean	SD	Median	Min	Max
K	6,426	395.93	28.50	396.22	324.82	529.37
1	2,648	480.10	17.70	479.26	402.92	568.29
2	8,549	488.04	18.06	488.05	410.52	572.22
3	9,760	496.78	19.91	497.55	401.70	572.83
4	10,528	506.69	19.27	507.71	416.66	580.67
5	7,739	517.86	16.77	518.18	446.72	589.91
6	7,099	524.73	18.90	524.73	447.15	599.41
7	5,048	530.25	20.92	530.12	444.07	599.41
8	7,788	535.19	21.75	535.35	455.49	599.41

Reliability

Internal-consistency reliability was estimated using an IRT model-based approach in which the variance of the conditional standard error of measurement (cSEM) for each student represents error variance, and the variance of the IRT theta scores represents total variance. Because reliability is the ratio of true score variance to total variance and true score variance equals 1 - error variance, the following model was used. Results are reported in Table 67. All reliability coefficients are in the mid to upper 0.90s indicating very strong measurement precision and reliability.

$$R_{xx} = \frac{(1 - \sum CSEM^2)}{Var(\theta)}$$

Table 67

IRT Model-Based Reliability Coefficients for AUTOreading by Grade

Grade	N	Reliability
K	7,215	0.97
1	2,651	0.94
2	8,576	0.94
3	9,808	0.95
4	10,574	0.95

5	7,743	0.95
6	7,110	0.96
7	5,073	0.97
8	7,838	0.97

Validity

Content Validity

The alignment of the skills assessed by AUTOreading subtests to the foundational reading skills of phonics, phonemic awareness, and fluency in the Big-5 reading domains provides evidence for content validity. A complete description of the subtests and how the skills are assessed is provided in the *FastBridge Assessments Content Description and Use Guide V5* (Illuminate Education, 2020).

Construct Validity

Construct validity is determined by the accumulation of evidence from multiple studies. For multiskill assessments like AUTOreading, one element of construct validity is the degree to which the subtest scores correlate. Generally, one would expect moderate to high correlations ranging from the 0.50s to 0.70s. Correlations below 0.50 are a threat to the validity of a composite based on a unidimensional model like the GRM model used to score AUTOreading. Correlations in the 0.80s or higher indicate redundancy across skills in the form of common method or construct variance or both.

Table 68 shows the correlation coefficient between subtest scores with each level of the AUTOreading test. Coefficients range from 0.60 to 0.80, with most in the upper 0.60s to lower 0.70s. These results provide good evidence that subtests measure distinct but related skills and that they can be combined into a single composite score. The correlation between each subtest and the overall composite score is shown in the first column. As expected, the correlations are very high because the variance of each subtest is included in the composite. The higher the correlation the greater the influence of a particular subtest on the overall score. The relatively strong influence of Letter Names in Level A is likely an artifact of using data from the fall of kindergarten. At that point, most students have underdeveloped letter-sound correspondence skills.

Table 68

Inter-Correlation by AUTOreading Subtest

	Composite	Letter Names	Letter Sounds	Encoding	Word ID	Vocabulary	Decoding	Matching Synonyms	Morphology
Level A									
Composite	1.00								
Letter Names	0.97	1.00							
Letter Sounds	0.85	0.72	1.00						
Level B									

Composite	1								
Encoding	0.94	-	-	1.00					
Word ID	0.93	-	-	0.80	1.00				
Vocabulary	0.89	-	-	0.76	0.77	1.00			
Level C									
Composite	1.00								
Word ID	0.92	-	-	-	1.00	-			
Decoding	0.86	-	-	-	0.74	-	1.00		
Matching Synonyms	0.86	-	-	-	0.71	-	0.68	1.00	
Morphology	0.79	-	-	-	0.60	-	0.60	0.60	1.00

Validity was evaluated using correlations with FastBridge adaptive reading assessment (aReading) and the FastBridge oral reading fluency assessment (CBMreading). aReading is a standards-based computer-adaptive assessment that measures general reading ability across the full continuum from pre-reader through advanced reading comprehender. It is an appropriate criterion measure because AUTOreading provides supplemental diagnostic information to support interpretation of students identified as at-risk based on aReading. Therefore, we hypothesized that AUTOreading scores would correlate moderately with aReading scores. We also chose CBMreading as another criterion measure because, in addition to serving as a skills-based diagnostic, AUTOreading is also used as a proxy for reading automaticity. Both AUTOreading and CBMreading use rate-based scoring to assess reading automaticity, but with different formats and content. Therefore, we hypothesized that AUTOreading scores would correlate moderately with CBMreading scores. Results were obtained from fall and winter universal screening administrations of these measures during the 2018–2019 school year and are reported in Table 69. The data include more than three-dozen school districts across the U.S. and across all urbanicity and free/reduced lunch strata.

As anticipated, the concurrent validity coefficients with aReading are moderate and generally in the 0.60s. The validity coefficients with CBMreading are strong and generally in the 0.70s. The strength of the relationship between AUTOreading scores and CBMreading words correct per minute score is evidence that AUTOreading scores approximate reading automaticity.

Table 69

AUTOreading Correlation with aReading and CBMreading.

	aReading		CBMreading			
Grade	N	Corr.	N	Corr.	Mean	SD
K	6,426	0.51	-	-	-	-
1	2,648	0.58	1,292	0.80	42.2	37.2
2	8,549	0.64	3,023	0.82	78.4	45.1
3	9,760	0.67	2,864	0.75	97.8	45.5
4	10,528	0.65	2,775	0.74	115.8	46.7

5	7,739	0.63	2,279	0.76	129.7	45.6
6	7,099	0.60	1,205	0.70	134.6	40.9
7	5,048	0.62	656	0.74	140.9	46.0
8–12	7,788	0.61	695	0.73	135.4	46.9

CBMmath Automaticity

Overview

CBMmath Automaticity is a suite of screening and progress monitoring curriculum-based measures (CBM) that evaluate the degree to which students have mastered and become fluent with basic math facts and operations. CBMmath Automaticity screening scores provide diagnostic information to support instructional planning, especially for students with poor computational and problem-solving skills. Results from CBMmath Automaticity supplement the aMath results and are included in the FastBridge Screening-to-Intervention report. This report guides instructional planning at the whole class, small group, and individual student level. AUTOreading also includes 20 alternate forms per test level and subskill level for frequent progress monitoring.

There are three leveled multiskill screening assessments:

- Level 1 assesses addition and subtraction facts through 20
- Level 2 assesses one- and two-digit addition, subtraction, and multiplication facts through 100
- Level 3 assesses two-digit addition, subtraction, multiplication, and division facts

Reliability

Alternate-form reliability was used to estimate the reliability and score consistency of CBMmath Automaticity measures. This is the most appropriate measure because the suite includes 20 alternate forms per level; it uses a timed, speeded approach to assess automaticity; and it is objectively scored by the computer algorithms. The data for the analysis are from actual progress monitoring scores collected during the 2018–2019 academic year and include data from more than 50 school districts across the U.S. representing all urbanicity and free/reduced lunch rate strata. Results are reported in Table 70. Coefficients are all near 0.80 and provide evidence of strong reliability.

Table 70

Alternate-Form Reliability of CBMmath Automaticity, by Level

Measure	Grade	N	Median	CI
Level 1	1–2	2,105	0.79	0.78 – 0.80
Level 2	2–4	8,535	0.78	0.78 – 0.78

Level 2	4–5	3,770	0.82	0.81 – 0.83

Test-retest reliability was also evaluated using repeat administration during a single screening period. Results are reported in Table 71. Coefficients are all near 0.80 and provide evidence of strong rank-order score stability.

Table 71

Test-Retest Reliability of CBMmath Automaticity, by Level

Measure	Grade	N	Median	CI
Level 1	1	98	0.79	0.76–0.88
Level 2	2–3	167	0.78	0.71–0.83
Level 2	4–5	86	0.82	0.72–0.87

Validity

Validity was evaluated using correlations with FastBridge adaptive math assessment (aMath). aMath is a standards-based computer-adaptive assessment that measures general math ability across the full continuum from early numeracy counting skills through advanced high school mathematics. It is an appropriate criterion measure because CBMmath automaticity provides supplemental diagnostic information to support interpretation of students identified as at-risk based on aMath. Therefore, we hypothesized that CBMmath automaticity scores would correlate moderately with aMath scores. The data comes from standard screening conducted in over 1,000 schools across the U.S. and representing all levels of urbanicity and free/reduced lunch rate strata.

Table 72 shows concurrent validity results and Table 73 shows predictive validity results. As expected, concurrent validity coefficients are moderate and somewhat stronger than predictive validity coefficients. A closer inspection of the data indicates a stronger relationship between the two assessments up to the point that students obtain automaticity, or about 60 correct per minute. This makes sense because once automaticity is obtained, greater rates should not improve math performance.

Table 72

Concurrent Validity of CBMmath Automaticity with aMath, by Level

Grade	Grade	N	Coefficient	CI Lower Bound	CI Upper Bound
Level 1	1	22,087	0.70	0.70	0.70
Level 2	2	53,322	0.65	0.65	0.65
Level 3	3	47,380	0.58	0.58	0.58

Table 73

Predictive Validity of CBMmath Automaticity with aMath, by Level

Grade	Grade	Season	N	Coefficient	CI Lower Bound	CI Upper Bound
Level 1	1	Fall	5,444	0.54	0.52	0.56
Level 1	1	Winter	6,203	0.59	0.57	0.61
Level 2	2	Fall	11,893	0.56	0.55	0.57
Level 2	2	Winter	16,102	0.60	0.59	0.61
Level 3	3	Fall	9,967	0.51	0.50	0.52
Level 3	3	Winter	14,293	0.54	0.53	0.55

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (2014). *Standards for Educational and Psychological Testing*. Washington DC: AERA Publications.

Brown-Chidsey, R. & Bickford, R. (2016). Practical handbook of multi-tiered systems of support: Building academic and behavioral success in schools. New York, NY: Guilford Press.

Browne, M. W., Cudeck, R., Bollen, K. A., & Long, J. S. (1993). Testing structural equation models. Sage Publications Inc.

Chafouleas, S.M., Sanetti, L.M., Kilgus, S.P., & Maggin, D.M. (2012). Evaluating sensitivity to behavioral change using Direct Behavior Rating Single-Item Scales. Council for Exceptional Children, 78(4), 491-505.

Christ, T.J., Riley-Tillman, T.C., & Chafouleas, S.M. (2009). Foundations for the development and use of Direct Behavior Rating (DBR) to assess and evaluate student behavior. Assessment for Effective Intervention, 34(4), 201–213.

Christ, T. J. & Aranas, Y. A. (2014). Best practices in problem analysis. *Best Practices in School Psychology V*, 2, 159–176.

Christ, T.J., Riley-Tillman, T.C. & Chafouleas, S.M. (2009). Foundations for the development and use of Direct Behavior Rating (DBR) to assess and evaluate student behavior. *Assessment for Effective Intervention*, 34(4), 201–213.

Deno, S. L. (1985). Curriculum-based measurement: The emerging alternative. *Exceptional children*, 52(3), 219–232.

Deno, S. L. (2005). Problem-solving assessment. In Brown-Chidsey, R. Assessment for intervention: A problem-solving approach (pp. 10–36). New York: Guilford. Glass, G. V. (1978). Standards and criteria. *Journal of Education*, 15(4), 237–261.

Dimitrov, D. M. (2003). Marginal true-score measures and reliability for binary items as a function of their IRT parameters. *Applied Psychological Measurement*, 27(6), 440–458.

Gignac, G. E. & Watkins, M. W. (2013). Bifactor modeling and the estimation of model-based reliability in the WAIS-IV. *Multivariate Behavioral Research*, 48, 639–662. https://doi.org/10.1080/00273171.2013.804398

Gresham, F. M. & Elliott, S. N. (2008). Social Skills Improvement System-Rating Scales. Pearson Assessments.

Hu, L.-T., & Bentler, P. M. (1995). Evaluating model fit. In R. H. Hoyle (Ed.), *Structural equation modeling: Concepts, issues, and applications* (pp. 76–99). Sage Publications, Inc.

Illuminate Education Inc., 2020. FastBridge Assessments Content Description and Use Guide

Illuminate Education Inc., 2020. Statistical Analysis of Bias of FastBridge Assessments

Jenkins, J. R. & Jewell, M. (1992). An examination of the concurrent validity of the Basic Academic Skills Samples (BASS). *Assessment for Effective Intervention*, 17(4), 273–288.

Kilgus, S. P., Chafouleas, S. M. & Riley-Tillman, T. C. (2013). Development and initial validation of the social and academic behavior risk screener for elementary grades. *School Psychology Quarterly*, 28, 210–226.

Kilgus, S. P. & Eklund, K. R. (2016). Consideration of base rates within universal screening for behavioral and emotional risk: A novel procedural framework. *School Psychology Forum: Research in Practice*, 10, 120–130.

Kilgus, S.P., Eklund, K., von der Embse, N.P., Taylor, C.N. & Sims, W.A. (2016). Psychometric defensibility of the Social, Academic, and Emotional Behavior Risk Screener (SAEBRS) Teacher Rating Scale and multiple gating procedure within elementary and middle school samples. *Journal of School Psychology*, 58, 21–39.

Kilgus, S.P., Eklund, K., von der Embse, N.P., Weist, M., Barber, A.J., Kaul, M. & Dodge, S. (2020). Structural validity and reliability of Social, Academic, Emotional Behavior Risk Screener-Student Rating Scale scores: A replication study.

Little, P. T. D. (2013). *Longitudinal structural equation modeling*. Guilford Press.

MacGinitie, W., MacGinitie, R., Maria, K., & Dreyer, L.G. (2000). Gates-MacGinitie Reading Tests. 4th ed. Itasca, IL: Riverside Publishing Company.

Markland, D. (2007). The golden rule is that there are no golden rules: A commentary on Paul Barrett's recommendations for reporting model fit in structural equation modelling. Personality and Individual Differences, 42(5), 851–858.

Morsy, L., Kieffer, M. & Snow, C. (2010). Measure for Measure: A Critical Consumers' Guide to Reading Comprehension Assessments for Adolescents. Final Report from Carnegie Corporation of New York's Council on Advancing Adolescent Literacy. *Carnegie Corporation of New York*.

National Center on Intensive Intervention (2020). Academic Screening Tools Chart Ratings Rubric. U.S. Office of Special Education Programs. Retrieved from https://intensiveintervention.org/sites/default/files/NCII_AcademicScreening_RatingRubric_2020-06-30.pdf.

Reise, S. P., Bonifay, W. E., & Haviland, M. G. (2013). Scoring and modeling psychological measures in the presence of multidimensionality. *Journal of Personality Assessment*, 95, 129–140. https://doi.org/10.1080/00223891.2012.725437

Salvia, J., Ysselydke, J. E. & Bolt, S. (2010). *Assessment in special and inclusive education* (11th ed.). Houghton Mifflin.

Shinn, M. R. (Ed.) (1989). *Curriculum-based measurement: Assessing special children*. Guilford Press.

Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. Psychometrika, 74, 107–120. https://doi.org/10.1007/s11336-008-9101-0

Silberglitt, B. & Hintze, J. (2005). Formative assessment using CBMreading cut scores to track progress toward success on state-mandated achievement tests: A comparison of methods. *Journal of Psychoeducational Assessment*, 23(4), 304–325.

Streiner, D. L. & Cairney, J. (2007). What's under the ROC? An introduction to receiver operating characteristic curves. *The Canadian Journal of Psychiatry/La Revue Canadienne de Psychiatrie*, 52, 121–128.

von der Embse, N.P., Kilgus, S.P., Iaccarino, S. & Levi-Nielson, S. (2013). Screening for student mental health risk: Diagnostic accuracy, measurement invariance, and predictive validity of Social, Academic, and Emotional Behavior Risk Screener-Student Rating Scale (SAEBRS-SRS). *School Mental Health*, 5(3), 119–129.

von der Embse, N.P., Iaccarino, S., Mankin, A., Kilgus, S.P. & Magen, E. (2016). Development and validation of the Social, Academic, and Emotional Behavior Risk Screener-Student Rating Scale. *Assessment for Effective Intervention*, 1–7.

Williams, K. T. (2004). GMADE Group mathematics assessment and diagnostic evaluation. Assessment and diagnostic evaluation. Technical manual.