mCLASS® Español

mCLASS Lectura
Technical Manual

Amplify.

Table of Contents

mCLASS Lectura Technical Manual

Introduction

mCLASS Lectura is designed to assess critical component skills in Spanish literacy development for Grades K–6 as identified by the National Literacy Panel for Language-Minority Children and Youth (August & Shanahan, 2006), including alphabet knowledge, phonological awareness, alphabetic understanding, decoding, fluency, and reading comprehension. As the grade level increases, foundational measures are phased out and measures of more complex skills are introduced in alignment with Spanish literacy development and instructional focus.

mCLASS Lectura takes a general outcome measurement (GOM) or curriculum–based measurement (CBM; see Deno, 1992) approach to reading assessment. In other words, the measures are designed to assess the most critical Spanish literacy skills using common assessment formats that follow an evidence-based trajectory of skill development, but are not tied to a specific Spanish literacy curriculum. mCLASS Lectura is designed to assess students' Spanish reading skills from the beginning of kindergarten through the end of sixth grade using a set of standardized measures that are brief and efficient to administer (each measure takes approximately 1–3 minutes to complete). Because mCLASS Lectura subtests are timed, fluency is considered as well as accuracy with the component skills. The subtests offered in specific grades are aligned to curriculum and instruction typical for each grade.

This manual is a compendium of technical information for two mCLASS Lectura studies that provide information about the composite score and subtests that contribute to the composite score. During 2020–2021, we conducted the mCLASS Lectura calibration study, which served to establish a scale of item difficulty that was then used to develop alternate, equivalent benchmark forms for use at the beginning of year (BOY), middle of year (MOY), and end of year (EOY) in Grades K–6. During 2021–2022, we conducted the mCLASS Lectura field study, which served to provide the data needed to develop a composite score, generate cut scores for student performance categories, and, subsequently, gather reliability and validity evidence of the composite and subtest scores by comparing student performance on other standardized assessments of Spanish literacy skills. This report will focus on the results from the two studies for Grades K–5. We begin by providing descriptions of the sample recruitment and selection procedures, the sample and measure descriptions, used within the studies. The results of the calibration study, including the development of the subtests, are presented subsequently. Next, we present the results for the mCLASS Lectura field study including subtest descriptives and correlations, reliability and validity evidence, and composite and cut score development. For Grade 6, additional study is planned to expand the sample size and build on the preliminary evidence of reliability, validity, and classification accuracy analyzed during the 2021–2022 school year.

Chapter 1: Sample Descriptions

Sample Recruitment and Selection Procedures

Amplify recruited elementary and middle schools from across the United States to participate in the mCLASS Lectura research during the 2020–2021 and 2021–2022 school years. Schools were recruited from mCLASS customers using the existing mCLASS Spanish assessment, Indicadores Dinámicos del Éxito en la Lectura (IDEL), through website postings and email contacts, as well as via connections to Amplify customer support managers and Amplify sales team members. Prior to reaching out to districts, the following criteria for participating schools were confirmed: a) students are enrolled in a dual-language program and/or are native Spanish speakers, b) participating students have a range of Spanish reading proficiency levels, and c) participating students must be enrolled in any of the target grade levels (K–6). Once eligibility criteria were met, information about the project, including participation requirements and incentives, were communicated to potential participating schools via a flier containing a link to a questionnaire schools were asked to complete if they were interested in discussing the study further. School staff then received a description of the study, selection criteria, and participation options. Amplify research staff reached out to all interested schools for a virtual meeting to discuss the research activities schools would be expected to complete by Time of Year (TOY). Schools were recruited until Amplify met or exceeded its recruitment goals or until it was no longer feasible for schools to assess students during the specified benchmark administration windows.

All students who were enrolled in a Dual Language Immersion (DLI) program and/or native Spanish speakers were eligible for participation and were included unless they would normally be excluded from typical assessments. At their discretion, schools could also opt not to assess students with disabilities who required assessment modifications.

Description of the mCLASS Lectura Research Samples

The mCLASS Lectura calibration study (2020–2021) consisted of one sample (Sample A), and the mCLASS Lectura field study (2021–2022) consisted of three samples (Samples B, C, and D). The three samples derived from the field study were used for specific purposes; Sample B, the largest sample composed of students with at least one mCLASS Lectura subtest, was used to generate normative information about mCLASS Lectura subtests, including descriptives and cut scores. Sample C, composed of students with at least one mCLASS Lectura subtest score and scores on the external criterion measures, was used as the reliability and validity sample. Sample D, composed of students with scores on all mCLASS Lectura subtests at each TOY and data on the external criterion measures, was used to examine the ability of mCLASS Lectura to accurately differentiate between students who were on track or at risk in other Spanish literacy assessments and to establish the cut scores for the mCLASS Lectura composite and subtests. Each sample was selected to answer specific research questions, which we articulate in more detail in the sections that follow.

Sample A, presented in Table 1, represents the mCLASS Lectura calibration study sample conducted during the 2020–2021 school year. Sample A was used to calibrate items to establish a scale of item difficulty, which was then used to develop alternate equivalent benchmark assessments of comparable test form difficulty for each subtest and grade level.

During the 2020–2021 mCLASS Lectura calibration study, data from Sample A was used to answer the following research questions for each TOY:

- What is the difficulty of each item within each subtest?
- What is the distribution of item difficulty within a subtest?
- To what degree are the items calibrated within the final benchmark forms so that we can establish a scale of item difficulty?

Table 1: mCLASS Lectura Calibration Study Sample A

Subtest	K	1	2	3	4	5
Fluidez en nombrar letras (FNL)	318	336				
Fluidez en la segmentación de sílabas (FSS)	276	325				
Fluidez en los sonidos de las letras (FSL)	336	338				
Fluidez en los sonidos de sílabas (LSS)	306	342				
Fluidez en la lectura de palabras (FEP)	223	327	301	225		
Fluidez en la lectura oral (FLO)		288	240	179	133	110
¿Cuál palabra? (CP)			180	145	147	121

Sample B, presented in Table 2, represents the norming sample and includes students with at least one mCLASS Lectura subtest score for at least one TOY during the 2021–2022 school year. This sample was used to compute the descriptive statistics for the mCLASS Lectura subtest scores. Sample B was selected to answer the following research questions:

- How do students perform on mCLASS Lectura subtests by TOY?
- What are the correlations between mCLASS Lectura subtests?

Three of the four census regions and five of nine census divisions were represented in Sample A (U.S. Department of Commerce, 2020). Students with diverse learning needs and representing multiple geographic regions (and dialectical variations of Spanish) were included in the national sample dataset. Students were enrolled in dual-language programs and/or were native Spanish speakers. Of the 32,933 students participating in the study, 50.4% were female, 48.1% were male, and 1.5% lacked data regarding gender. With respect to race/ethnicity, 63.2% of students were Hispanic/Latino, 14.1% were White, 4.4% were Black/African American, 1.5% were two or more races, and 1.0% identified as American Indian/Alaska Native, Asian, or Native Hawaiian/Pacific Islander; race/ethnicity data were unavailable for 15.8% of the sample. In addition, 30.0% of the students in the sample were considered English Learners (ELs), 52.6% of the students identified English as their primary language, and approximately 17.4% of the sample lacked language data. Finally, 3.6% of the students were eligible for special education, 50.8% were not eligible, and 45.5% of the sample lacked data regarding special education eligibility.

Table 2: Sample B Demographic Characteristics by Grade Level

		All	K	1	2	3	4	5
Sample Size								
District		56	52	53	47	36	11	10
Schools		372	274	291	274	159	82	60
Students	n	32933	8538	8499	7902	4124	2209	1661
	%	100	25.9	25.8	24.0	12.5	6.7	5.0
Gender								
Female	n	16612	4296	4333	3993	2073	1080	837
	%	50.4	50.3	51.0	50.5	50.3	48.9	50.4
Male	n	15826	4112	4060	3776	1997	1081	800
	%	48.1	48.2	47.8	47.8	48.4	48.9	48.2
Missing	n	495	130	106	133	54	48	24
	%	1.5	1.5	1.2	1.7	1.3	2.2	1.4
Ethnicity								
Alaska Native & American Indian	n	35	6	10	6	5	7	1
	%	0.1	0.1	0.1	0.1	0.1	0.3	0.1
Asian	n	218	44	60	69	32	7	6
	%	0.7	0.5	0.7	0.9	0.8	0.3	0.4
Black	n	1461	436	373	363	215	40	34
	%	4.4	5.1	4.4	4.6	5.2	1.8	2.0
Hispanic-Latino	n	20829	5063	5060	4666	2878	1827	1335
	%	63.2	59.3	59.5	59.0	69.8	82.7	80.4

		All	K	1	2	3	4	5
Ethnicity								
Multiracial	n	478	162	132	93	65	14	12
	%	1.5	1.9	1.6	1.2	1.6	0.6	0.7
Native Hawaiian or Other Pacific Islander	n	62	15	17	11	12	5	2
	%	0.2	0.2	0.2	0.1	0.3	0.2	0.1
White	n	4659	1266	1148	1052	801	197	195
	%	14.1	14.8	13.5	13.3	19.4	8.9	11.7
Missing	n	5191	1546	1699	1642	116	112	76
	%	15.8	18.1	20.0	20.8	2.8	5.1	4.6
English Learner Status								
Yes	n	9883	2530	2476	2265	1313	801	498
	%	30.0	29.6	29.1	28.7	31.8	36.3	30.0
No	n	17311	4373	4242	3919	2599	1172	1006
	%	52.6	51.2	49.9	49.6	63.0	53.1	60.6
Missing	n	5739	1635	1781	1718	212	236	157
	%	17.4	19.1	21.0	21.7	5.1	10.7	9.5
Special Education								
Yes	n	1197	331	315	271	129	91	60
	%	3.6	3.9	3.7	3.4	3.1	4.1	3.6
No	n	16739	3720	3908	3950	2228	1662	1271
	%	50.8	43.6	46.0	50.0	54.0	75.2	76.5
Missing	n	14997	4487	4276	3681	1767	456	330
	%	45.5	52.6	50.3	46.6	42.8	20.6	19.9

Sample C, presented in Table 3, represents a subset of Sample B. Sample C consists of students with at least one mCLASS Lectura subtest score and the following criterion measures described at each TOY during the 2021–2022 school year. Sample C was used to evaluate the reliability and validity of the mCLASS Lectura Composite Score and subtest scores. Sample C was selected to answer the following research questions:

- What are the psychometric properties of the mCLASS Lectura Composite Score and subtests?
- What is the reliability of the mCLASS Lectura Composite Score and subtests?
- How well do the mCLASS Lectura Composite Score and subtests correlate with an external criterion measure?

Three of the four census regions and six of the nine census divisions were represented in Sample B (U.S. Department of Commerce, 2020). Of the 3,995 students in Sample C with at least one mCLASS Lectura subtest score and scores on the external criterion measures, 49.3% were female, 46.2% were male, and approximately 4.5% lacked data regarding gender. With respect to race/ethnicity, 62.0% of students were Hispanic/Latino, 17.1% were White, 2.2% were Black/African American, 1.4% were two or more races, and 1.1% identified as American Indian/Alaska Native, Asian, or Native Hawaiian/Pacific Islander; race/ethnicity data were unavailable for approximately 16.2% of the sample. In addition, 19.5% of the students in the sample were considered ELs, 46.1% of the students identified English as their primary language, and approximately 34.4% of the sample lacked language data. Finally, 3.6% of the students were eligible for special education, 44.8% were not eligible, and 51.6% of the sample lacked data regarding special education eligibility.

Table 3: Sample C Demographic Characteristics by Grade Level

		All	K	1	2	3	4	5
Sample Size								
District		15	11	14	14	8	8	8
Schools		32	13	25	23	11	13	11
Students	n	3995	408	1140	865	556	540	486
	%	100	10.2	28.5	21.7	13.9	13.5	12.2
Gender								
Female	n	1969	596	413	265	260	226	837
	%	209	52.3	47.7	47.7	48.1	46.5	50.4
Male	n	596	537	400	240	235	239	800
	%	413	47.1	46.2	43.2	43.5	49.2	48.2

		All	K	1	2	3	4	5
Gender								
Missing	n	226	4	7	52	51	45	21
	%	49.3	1.0	0.6	6.0	9.2	8.3	4.3
Ethnicity								
Alaska Native & American Indian	n	52.3	0	2	2	2	7	1
	%	47.7	0.0	0.2	0.2	0.4	1.3	0.2
Asian	n	47.7	0	6	8	4	3	3
	%	48.1	0.0	0.5	0.9	0.7	0.6	0.6
Black	n	46.5	0	32	27	11	8	11
	%	1846	0.0	2.8	3.1	2.0	1.5	2.3
Hispanic-Latino	n	195	261	812	542	303	292	267
	%	537	64.0	71.2	62.7	54.5	54.1	54.9
Multiracial	n	400	4	16	4	8	11	12
	%	240	1.0	1.4	0.5	1.4	2.0	2.5
Native Hawaiian or Other Pacific Islander	n	235	0	2	1	0	0	0
	%	239	0.0	0.2	0.1	0.0	0.0	0.0
White	n	46.2	74	111	105	134	125	136
	%	47.8	18.1	9.7	12.1	24.1	23.1	28.0
Missing	n	47.1	69	159	176	94	94	56
	%	46.2	16.9	13.9	20.3	16.9	17.4	11.5
English Learner Status								
Yes	n	43.5	33	226	154	138	121	109
	%	49.2	8.1	19.8	17.8	24.8	22.4	22.4

		All	K	1	2	3	4	5
English Learner Status								
No	n	1841	227	574	367	220	216	237
	%	46.1	55.6	50.4	42.4	39.6	40.0	48.8
Missing	n	1373	148	340	344	198	203	140
	%	34.4	36.3	29.8	39.8	35.6	37.6	28.8
Special Education								
Yes	n	145	19	41	29	24	14	18
	%	3.6	4.7	3.6	3.4	4.3	2.6	3.7
No	n	1789	150	585	403	205	218	228
	%	44.8	36.8	51.3	46.6	36.9	40.4	46.9
Missing	n	2061	239	514	433	327	308	240
	%	51.6	58.6	45.1	50.1	58.8	57.0	49.4

Sample D, presented in Table 4, represents a subset of Samples B and C and consists of students with complete data on the mCLASS Lectura subtests used for computing the composite score at each TOY and the EOY criterion measure during the 2021–2022 school year. This sample was used to examine the classification accuracy of mCLASS Lectura for accurately differentiating between students who were on track or at risk on other standardized assessments of Spanish literacy and to establish the cut scores for the mCLASS Lectura Composite Score and subtest scores. Sample D was selected to answer the following research questions:

- What is the classification accuracy of the mCLASS Lectura composite and subtests cut scores?
- How well does the mCLASS Lectura Composite Score and subtest scores predict performance on an external criterion measure?

Three of the four census regions and six of the nine census divisions were represented in Sample D (U.S. Department of Commerce, 2020). Of the 1,864 students contributing to Sample D, 52.4% were female, 47.5% were male, and less than 1% lacked data regarding gender. With respect to race/ethnicity, 69.5% of students were Hispanic/Latino, 22.5% were White, 1.9% were Black/African American, 2.1% were two or more races, and 1.3% identified as American Indian/Alaskan Native, Asian, or Native Hawaiian/Pacific Islander; race/ethnicity data were unavailable for approximately 2.7% of the sample. Approximately 44.2% of the students were eligible to receive free or reduced lunch, 51.8% were not eligible, and data were unavailable for 4.0% of the sample. In addition, 25.4% of the students in the sample were considered ELs, 71.9% of the students identified English as their primary language, and language data were unavailable for approximately 2.7% of the sample. Finally, 46.3% of the students' primary language at home was Spanish,

42.4% of students' primary home language was English, 7.6% of students spoke a language other than Spanish or English at home, and 3.6% of the sample lacked language data. Although instructional models used by the participating Districts varied, students in Grades K–2 were receiving the majority of their literacy instruction (80% or more) in Spanish while students in Grades 3–5 were receiving approximately half of their literacy instruction (50%) in Spanish.

Table 4: Sample D Demographic Characteristics by Grade Level

		All	K	1	2	3	4	5
Sample Size								
District		11	6	10	8	6	5	5
Schools		22	7	18	11	8	7	6
Students	n	1864	181	537	333	310	248	255
	%	100	9.7	28.8	17.9	16.6	13.3	13.7
Gender								
Female	n	977	94	281	182	165	127	128
	%	52.4	51.9	52.3	54.7	53.2	51.2	50.2
Male	n	886	86	256	151	145	121	127
	%	47.5	47.5	47.7	45.3	46.8	48.8	49.8
NA	n	1	1	0	0	0	0	0
	%	0.1	0.6	0.0	0.0	0.0	0.0	0.0
Ethnicity								
Alaska Native & American Indian	n	10	0	1	0	2	6	1
	%	0.5	0.0	0.2	0.0	0.6	2.4	0.4
Asian	n	11	0	3	5	0	1	2
	%	0.6	0.0	0.6	1.5	0.0	0.4	0.8
Black	n	35	0	7	8	9	6	5
	%	1.9	0.0	1.3	2.4	2.9	2.4	2.0

		All	K	1	2	3	4	5
Ethnicity								
Hispanic-Latino	n	1295	154	422	227	186	151	155
	%	69.5	85.1	78.6	68.2	60.0	60.9	60.8
Multiracial	n	39	2	11	4	6	6	10
	%	2.1	1.1	2.0	1.2	1.9	2.4	3.9
Native Hawaiian or Other Pacific Islander	n	3	0	2	1	0	0	0
	%	0.2	0.0	0.4	0.3	0.0	0.0	0.0
White	n	420	24	83	63	95	75	80
	%	22.5	13.3	15.5	18.9	30.6	30.2	31.4
Missing	n	51	1	8	25	12	3	2
	%	2.7	0.6	1.5	7.5	3.9	1.2	0.8
Free & Reduced Lunch Status								
FRL Eligible	n	823	104	303	158	104	69	85
	%	44.2	57.5	56.4	47.4	33.5	27.8	33.3
FRL Not Eligible	n	967	64	218	161	181	177	166
	%	51.9	35.4	40.6	48.3	58.4	71.4	65.1
Missing	n	74	13	16	14	25	2	4
	%	4.0	7.2	3.0	4.2	8.1	0.8	1.6
English Learner Status								
Yes	n	474	84	185	121	32	28	24
	%	25.4	46.4	34.5	36.3	10.3	11.3	9.4
No	n	1340	96	348	198	253	218	227
	%	71.9	53.0	64.8	59.5	81.6	87.9	89.0

		All	K	1	2	3	4	5
Home Language								
Missing	n	50	1	4	14	25	2	4
	%	2.7	0.6	0.7	4.2	8.1	0.8	1.6
Spanish	n	863	85	285	165	116	106	106
	%	46.3	47.0	53.1	49.5	37.4	42.7	41.6
English	n	791	32	184	147	164	129	135
	%	42.4	17.7	34.3	44.1	52.9	52.0	52.9
Other	n	142	53	56	7	5	11	10
	%	7.6	29.3	10.4	2.1	1.6	4.4	3.9
Missing	n	68	11	12	14	25	2	4
	%	3.6	6.1	2.2	4.2	8.1	0.8	1.6

Research Procedures

mCLASS Lectura Calibration Study

During the 2020–2021 school year, data for the mCLASS Lectura calibration study were collected by Amplify data collectors. Data collectors attended a half-day, web-based training on standardized administration and scoring procedures for the mCLASS Lectura subtests prior to the administration of the measures. Reliability checks for all subtests were completed by all trained staff before assessing students to confirm their ability to administer and score the assessments according to standardized procedures.

All assessors administered all mCLASS Lectura subtests to each student individually across kindergarten through Grade 6. For each of the subtests—FNL, FSL, LSS, FSS, and FEP—four forms were developed. Common items on adjacent forms were designed so that all items could be placed on the same scale for item analysis. Students were randomly assigned to one of the four forms and were given sufficient time to respond to all items. All items within each subtest were administered to all students. In addition, nine FLO passages and six ¿Cuál Palabra? (CP) passages were also developed by grade level. Students were randomly assigned to one of 10 packets, each of which consisted of three passages. Students were required to read each passage in its entirety before moving on to the next passage. CP was group-administered via paper and pencil in each participating classroom. Students were randomly assigned to one of 10 packets that included three passages counterbalanced to minimize order effects. Assessors distributed a packet to each student, read the instructions for completing the subtest, and asked them to begin. Students were asked to turn their packet over (face-down) once they had finished responding to all of the passages.

Field Study

During the 2021–2022 school year, all assessments were administered to students by teachers, reading coaches, and Amplify data collectors. All teachers, reading coaches, and data collectors administering the assessments attended a half-day, web-based training on standardized administration and scoring procedures for the mCLASS Lectura subtests and all external measures prior to the opening of each benchmark assessment window. Assessment windows were set by each district and assessments were administered at BOY, MOY, and EOY. Reliability checks for all measures were completed by all trained staff before assessing students to confirm their ability to administer and score the assessments according to standardized procedures.

All assessors administered all mCLASS Lectura subtests to each student individually across kindergarten through Grade 6 during each benchmark window, with the exception of CP. CP was group-administered to all students in classrooms or the school's computer lab following standard procedures. CP was administered via paper and pencil at BOY and online at MOY and EOY.

Students were administered multiple external criterion assessments of Spanish literacy to establish evidence of concurrent and predictive validity (that is, the degree to which mCLASS Lectura results correlate with results from measures that have been previously validated for providing accurate information about students' Spanish literacy skills). Students in kindergarten were administered the Análisis de Palabras subtest of Batería IV Woodcock-Muñoz, a parallel Spanish version of the Woodcock-Johnson IV (WM-AP; Woodcock et al., 2017); students in Grades 1 through 3 were administered Star Assessments for Spanish - Early Literacy (SELSp, Renaissance, 2021); and students in Grades 4 through 6 were administered Star Assessments for Spanish - Reading (SRSp, Renaissance, 2018). WM-AP, SELSp, and SRSp were administered at BOY, MOY, and EOY, and within 2 weeks of the administration of mCLASS Lectura.

Measure Descriptions

mCLASS Lectura - Brief Description

mCLASS Lectura is a collection of subtests designed to be administered at each benchmark period (i.e., BOY, MOY, and EOY) in kindergarten through sixth grade. These subtests assess students' alphabet knowledge, phonological awareness, alphabetic understanding, decoding, fluency, and reading comprehension skills. As the grade level increases, foundational measures (e.g., measures of phonological awareness and alphabetic understanding) are phased out and measures of more complex skills (e.g., fluency and reading comprehension) are introduced in alignment with Spanish literacy development and instructional focus. All measures for each respective grade level are administered during each benchmark period. See Table 5 for a brief description of each subtest. For more detailed information about the mCLASS Lectura subtests, please see the *mCLASS Lectura Administration and Scoring Guide*.

Table 5: Skills Assessed by mCLASS Lectura Subtest

Skill	Subtest	Grades	Description
Alphabet Knowledge	Fluidez en nombrar letras (FNL) – *Letter Naming Fluency*	K–1	Students identify the names of randomly mixed uppercase and lowercase letters on a printed stimulus form. The measure score is the number of letters correctly named in 1 minute.
Phonological Awareness	Fluidez en la segmentación de sílabas (FSS) – *Syllable Segmentation Fluency*	K–1	Students segment orally spoken words into syllables. The measure score is the total number of syllables correctly produced in 1 minute.
Phonics and Alphabetic Principle	Fluidez en los sonidos de las letras (FSL) – *Letter Sound Fluency*	K–1	Students identify the sounds made by randomly mixed uppercase and lowercase letters on a printed stimulus form. The measure score is the number of correct letter sounds produced in 1 minute.
Phonics and Alphabetic Understanding	Fluidez en los sonidos de sílabas (LSS) – *Syllable Sound Fluency*	K–1	Students decode orthographically regular syllables composed of two, three, or four phonemes (e.g., VC, CVC, CCV, CVCC, CVVC) on a printed stimulus form. The measure score is the number of syllables read correctly in 1 minute.
Phonics and Decoding	Fluidez en las palabras (FEP) – *Word Reading Fluency*	K–3	Students are given a printed stimulus form containing real words of increasing complexity, out of context. They read as many whole words (blending all sounds) as they can. The measure score is the number of whole words read correctly in 1 minute.
Fluency	Fluidez en la lectura oral (FLO) – *Oral Reading Fluency*	1–6	Students read aloud printed grade-level passages of authentically written Spanish text. The measure scores are the number of words read correctly in 1 minute and the percentage of words read accurately.
Comprehension	¿Cuál palabra? (CP) – *Maze*	2–6	In this group-administered maze measure, students are given a printed or online reading passage in which approximately every seventh word is replaced by a multiple choice box that includes the original word and two distractors. The students read the passage silently and select the word in each box that best fits the meaning of the sentence. The measure score is one-half the number of incorrect responses subtracted from the number of correct responses selected in 3 minutes.

External Criterion Measures

Woodcock Muñoz IV Análisis de Palabras (WM-AP): WM-AP assesses students' ability to apply phonic and structural analysis skills to orthographically regular Spanish nonsense and/or low-frequency words. Students in kindergarten through sixth grade are administered a maximum of 34 items. The initial items require students to produce the sounds of individual letters. The remaining items require students to

read aloud letter combinations and syllables that follow orthographic, morphological, and lexical rules of Spanish but in the context of nonsense or low-frequency words. The items become more difficult and the complexity of the nonsense words increases. WM-AP has a median reliability of .91–.96 in the 5–12 age range (Woodcock et al., 2017). WM-AP served as the kindergarten external criterion measure for mCLASS Lectura.

Star Early Literacy Spanish (SELSp): SELSp is a computer-adaptive assessment designed to measure the early literacy skills of beginning Spanish readers in two broad domains: Word Knowledge and Skills, and Comprehension Strategies and Constructing Meaning (Renaissance Learning, 2021). These broad domains include 10 subdomains assessing the following skills: Visual Discrimination, Concept of Word, Phonemic Awareness, Alphabetic Principle, Phonics, Structural Analysis, Vocabulary, Sentence-Level Comprehension, Paragraph-Level Comprehension, and Accentuation (Renaissance Learning, 2021). In this computer-adaptive assessment, each administration consists of 27 items of varying difficulty based on the student's responses presented in multiple choice format (three answer choices per item). Each item consists of a combination of audio instructions, an on-screen prompt in the form of a cloze stem containing text or graphics, and three answer choices containing letters, words, graphics, or numbers. SELSp takes approximately 10–15 minutes for students to complete. Similar to mCLASS Lectura, it is intended as a screening and progress monitoring assessment to track student progress and instructional needs.

In Grades 1–3, scaled score generic reliability for SELSp ranges from 0.83–0.88; split-half reliability ranges from 0.81–-0.87; and alternate form reliability ranges from 0.73–0.75. Concurrent validity with two Spanish easyCBM subtests for Grade 2 ranges from 0.67–0.72 (Renaissance Learning, 2021).

SELSp total scaled scores were used in the present analysis rather than scores from the seven subscales within SELSp because students may only see a limited number of items in some domains based on their item responses. Thus, scaled scores are considered the strongest estimate of a student's overall reading skills at a particular time (Renaissance Learning, 2014). SELSp served as the Grades 1–-3 external criterion measure for mCLASS Lectura.

Star Reading Spanish (SRSp): SRSp is a computer-adaptive assessment designed to measure Spanish reading achievement in five content domains: Word Knowledge and Skills, Comprehension Strategies and Constructing Meaning, Analyzing Literary Text, Understanding Author's Craft, and Analyzing Argument and Evaluating Text (Renaissance Learning, 2021). Students are administered 34 items of varying difficulty based on the student's responses that measure reading comprehension. SRSp takes approximately 20 minutes for students to complete. It is intended to provide data on students' reading skills so educators can set goals, respond quickly to student needs, monitor progress, and maximize growth.

The SRSp technical manual reports reliability coefficient ranges for Grades 1 through 6 as follows: split-half reliability of 0.87–0.94; alternate form reliability of 0.69–0.82; and generic reliability (i.e., calculated from the conditional error variance of item response theory (IRT) ability estimates) of 0.91–0.95 (Renaissance Learning, 2021). Concurrent validity coefficients for Grades 2 through 5 ranged from 0.58 to 0.68 with the State of Texas Assessments of Academic Readiness Standards Test (STAAR) and Spanish easyCBM subtests (Renaissance Learning, 2021).

SRSp total scaled scores were used in the present analysis rather than scores from the five subscales within SRSp because students may only see a limited number of items in some domains based on their item responses. Thus, scaled scores are considered the strongest estimate of a student's overall reading skills at a particular time (Renaissance Learning, 2014). SRSp served as the Grades 4–5 external criterion measure for mCLASS Lectura.

Development of mCLASS Lectura Subtests

Assessment Design, Development, and Item Calibration

Across all measures, alternate forms included within a grade level are designed to be of equivalent difficulty to allow for measurement of growth in student performance over time. With alternate, equivalent forms we can reasonably infer that changes in students' scores are due to increases in student skill and not differences in the difficulty of the content. Within each alternate form of the mCLASS Lectura subtests that include word lists (e.g., FSS and FEP), items become increasingly difficult based on frequency in print, age of acquisition, and frequency of morphological patterns. First we provide a description about considerations for the selection of words (items) for word-based measures. Then we provide a brief description of the systematic process used to create benchmark forms for administration at BOY, MOY, and EOY for each grade level. In addition, we provide empirical evidence of form equivalence based on the mCLASS Lectura calibration study; item difficulties for each item by TOY obtained from the item calibration study are available upon request. We engaged in the same assessment development processes described in the paragraphs that follow to develop progress monitoring forms for each of the subtests using larger item pools to minimize the repetition of any given item (i.e., an orthographically regular Spanish syllable or word) across forms. Twenty alternate forms have been developed for FSL, FSS, LSS, FEP, and FLO, and 10 alternate forms have been developed for CP. During the 2022–2023 school year, data will be collected to gather empirical evidence for progress monitoring subtests and determine which subtests are most sensitive to growth.

Care was taken during the development of mCLASS Lectura syllable-based measures (i.e., LSS) to ensure that the syllables selected for inclusion in the forms took into consideration the frequency of letter sounds as well as making sure that the syllables used did not violate any grammatical, morphological, or spelling rules in Spanish. Similarly, item development for mCLASS Lectura word-based measures (i.e., FSS and FEP) began with a systematically and strategically gathered corpus of words to be used across all subtests. Words were also selected from language assessments designed for use with Spanish-English bilinguals in the U.S. (i.e., the Expressive One Word Picture Vocabulary Test, Brownell, 2012; MacArthur-Bates Communication Development Inventories in Spanish, Jackson Maldonado et al., 2003); Spanish reading curricula used in the U.S. (i.e., Spanish language arts curriculum: Maravillas Wonders, n.d.); and other sources such as the frequency of words in Spanish (Davies & Davies, 2018). The words selected for each grade level were directly related to their designated age level on diagnostic assessments (i.e., 5–6 years old = Kindergarten versus 7–8 years old = first grade) and in what grade they were targeted in our curriculum review.

Cross-cultural influences were also taken into consideration. Culture influences what people talk about, which in turn influences the types and frequency of words that are learned early on in language development (Hammer & Rodriguez, 2012; Peña et al., 2012). Research indicates that there is significant variation between Spanish-speaking cultures represented across the U.S. Consequently, per recommendations from research (Peña et al., 2012), we took into consideration (a) the difficulty of items, (b) the lexical frequency, and (c) the syntactic structures that may highlight one-word type over another during item selection and form development. These considerations also apply to the passages developed for connected text reading (i.e., FLO and CP), in that familiarity with the content and semantic choices such as the vocabulary selected can influence the difficulty of the items and student motivation and interest (August & Shanahan, 2006).

We also incorporated reviews from a variety of stakeholders, including academic biliteracy experts and

educators in various levels (e.g., superintendents and classroom teachers) from multiple regions of the country, during our iterative item development process to ensure we were representative of not only Spanish spoken on the mainland in the U.S., but also in Puerto Rico.

During the 2020–2021 school year, to help construct mCLASS Lectura subtests, we conducted an item calibration study that examined the qualities of the items in the item pool so we could select items for constructing the benchmark forms. We conducted item analyses using classical test theory (CTT) and IRT approaches. With respect to CTT analyses, we examined item difficulty, or the proportion of students who answered an item correctly, as well as item differentiates between high-performing and low-performing students (as measured by point-biserial correlations). We also examined item difficulty and item fit statistics using the Rasch model, which is a probabilistic model based upon a latent trait (i.e., reading proficiency) that allows for conjoint measurement of persons and items on the same scale. The Rasch model assumes that the probability of a given item-person interaction is governed by the difficulty of the item and the person's ability; using an IRT approach, the item difficulty represents the location on the latent trait scale at which the probability of a correct response is equal to the probability of an incorrect response (0.50) and, subsequently, that students whose ability level on the latent trait scale is greater than the item difficulty will have a higher probability of responding to the item correctly. We identified items with extreme difficulties and very low (i.e., < 0.2) or negative point-biserial correlations and high item fit statistics (i.e., > 1.5) and investigated them for potential causes. If the unusual statistics were not caused by some simple errors (e.g., miscoding of response), we either removed or revised the items until all items were of good quality. In the sections that follow, we describe in more detail the development of each mCLASS Lectura subtest and present empirical evidence of form equivalence for the BOY, MOY, and EOY benchmark forms for the subtest at the appropriate grade levels.

Fluidez en nombrar letras (FNL; Letter Names) and Fluidez en los sonidos de las letras (FSL; Basic Phonics)

During FNL and FSL administration, students are presented with a page of 100 uppercase and lowercase letters; for FNL, they are asked to name as many letters as they can in 1 minute, while for FSL they are asked to identify as many letter sounds as they can in 1 minute. To create the equivalent alternate FNL and FSL forms for kindergarten and Grade 1, we engaged in the following steps: First, we sorted the list of letters by case (lowercase then uppercase) and from smallest to largest (by the standard error associated with the Rasch item difficulties). Second, we used this information to identify which letters/items had the most stable item difficulty estimate (i.e., lowest standard error) to retain. At this step, we also chose to remove LL and RR from the list of viable items because neither of these digraphs appear at the beginning of a word in Spanish. Third, we sorted the pool of letters by Rasch item difficulty from smallest to largest (each letter appears only once in the item pool). Next, we divided the item pool into three blocks, with two blocks of 20 letters each and one block of 15 letters each, and then sorted the letters within each block for each benchmark form. This resulted in three alternate forms for a grade level that have, for example, the same 20 letters at the beginning of the form but in different orders. Finally, and as a result of this systematic process, the first 54 letters (after dropping LL and RR from the item pool) in each form are unique, and care was taken to randomly select items from the remaining items (i.e., not in Blocks 1 and 2) from sets of items still sorted by item difficulty such that the items in each block have similar average difficulty and require the same total number of syllables to produce the names of the letters within each row across the alternate forms. We also took care to ensure that neither the same items nor letters that are close to each other in the alphabet (e.g., *m*, *n*, *ñ*) appear in proximity within each block of items on the form. Taking all these form development constraints and Rasch item difficulties into consideration, we were able to generate three alternate equivalent forms for kindergarten and Grade 1 with the corresponding form difficulties (i.e.,

average Rasch item difficulties for the form). Table 6 shows that the FNL form difficulties were around -3.39 and -2.13 for Grade K and Grade 1, and the FSL form difficulties were around -3.14 and -2.18 for Grade K and Grade 1.

Table 6: Evidence of Form Equivalence for FNL and FSL by Grade Level and TOY

	FNL			FSL		
Grade	**BOY**	**MOY**	**EOY**	**BOY**	**MOY**	**EOY**
K	–3.388	–3.388	–3.388	–3.138	–3.138	–3.138
1	–2.139	–2.129	–2.133	–2.178	–2.176	–2.178

Fluidez en la segmentación de sílabas (FSS; Phonological/Syllable Awareness)

During this phonological awareness task, the examiner orally presents one word at a time to the student and the student is asked to segment the word into as many syllables as they can. Words for this measure were derived from the carefully constructed word bank described previously. To create alternate, equivalent forms for kindergarten (40 items each) and Grade 1 (50 items each), we engaged in a systematic, multi-step process. First, we built item pools of anchor items—to allow for horizontal equating of forms within grade and vertical equating of forms across grades—and unique items, and we ordered each set of items by their Rasch item difficulty from least to greatest item difficulty. The kindergarten item bank contained 17 anchor items and 84 unique items, and the Grade 1 item bank contained (the same) 17 anchor items and 81 unique items. From the bank of potential anchor items, 15 words were selected based on the scatter plots of item difficulties for kindergarten and Grade 1. Then we divided the unique items for each grade level into blocks (16–17 items per block for kindergarten and 11–12 items per block for Grade 1) and the 15 anchor items into three blocks (five items each) so the average difficulty of each item block could be calculated and compared. Blocks of items were ordered by their average difficulty, the blocks of unique and anchor items were finalized, and blocks of items equivalent in both difficulty and total number of syllables were identified for each of the three benchmark forms. Finally, the 15 anchor items were divided into 3 blocks, comprising five randomly selected anchor items from each block, and those anchor item blocks were placed in the same rows in the first half of each benchmark form (to increase the likelihood that students would respond to the anchor items). As a result of this systematic process, words within each FSS form are ordered from easiest to most challenging based on empirical item difficulty, not based on the number of syllables, letters, or sounds within the word. For example, although the word *húmedo* contains three syllables (/hú/ /me/ /do/) and is longer than the word *feliz* (/fe/ /liz/), *húmedo* appears before *feliz* in the Grade 1 benchmark forms because it is easier than *feliz* (item difficulties of -2.09 and -1.42, respectively). Collectively this approach was used to generate three forms of 40 items each for administration in kindergarten and three forms of 50 items each for administration in Grade 1. We then verified that each benchmark form within kindergarten and Grade 1 was of equivalent difficulty (i.e., FSS form difficulties were around -1.51 for Grade K and -1.54 for Grade 1) and contained the same number of syllables (i.e., 146 syllables at each TOY; see Table 7).

Table 7: Evidence of Form Equivalence for FSS by Grade Level and TOY

	BOY		MOY		EOY	
Grade	Form Difficulty	Number of Syllables	Form Difficulty	Number of Syllables	Form Difficulty	Number of Syllables
K	–1.51	110	–.50	110	–1.51	110
1	–.54	146	–1.53	146	–1.54	146

Fluidez en los sonidos de sílabas (LSS; Beginning Decoding)

In this task, students are presented with a page of printed orthographically regular Spanish syllables (50 for kindergarten and 60 for Grade 1) and asked to read as many syllables as they can in 1 minute. To create three alternate, equivalent forms for administration in kindergarten and Grade 1, we engaged in the following steps: First we established the constraints of the forms for each grade level, including: (1) the number of items (50 and 60 for kindergarten and Grade 1, respectively), (2) the number of items per row and rows per form (5 items per row, totaling 10 rows per form for kindergarten and 12 rows per form for Grade 1), and (3) the number of anchor items required for horizontal and vertical equating (15 anchors, organized in three intact rows, ordered by item difficulty). Once anchor items across the range of item difficulties and number of sounds per syllable were identified from the item pool, they were excluded from the remaining steps of manipulating and ordering the item pool. The remaining unique items for each grade (n = 82 for kindergarten and n = 84 for Grade 1) were ordered with respect to their Rasch item difficulties from least to greatest and then divided into blocks of nine or 10 items to randomly sample from to ensure that row contained items that (a) were comparable in their average item difficulty across the row (i.e., an absolute difference in difficulty of .012) and (b) contained the same number of sounds in each row across forms. Finally, rows of items were ordered within forms by average difficulty. Table 8 shows that for both grades, the difficulties of the three benchmark forms were similar (i.e., the form difficulties in kindergarten were around -0.78 and the form difficulties in Grade 1 were around -2.03).

Table 8: Evidence of Form Equivalence for LSS by Grade Level and TOY

Grade	BOY	MOY	EOY
K	–0.784	–0.781	–0.783
1	–2.026	–2.030	–2.027

Fluidez en las palabras (FEP; Beginning/Advanced Decoding)

In this task, students in Grades K–3 are presented with a page of 100 real Spanish words (pulled from the corpus of words described at the beginning of this section) and asked to read as many words as they can in 1 minute. Monosyllabic and multisyllabic words with syllable structures of varying complexity constitute the 100 words on each form. To construct alternate, equivalent forms using strategically selected anchor items to allow for horizontal and vertical equating, we began by identifying the number of anchor items needed

for both types of equating. For kindergarten and Grade 3, 16 anchor items were selected, leaving 84 unique items in the unique item pool. For Grades 1 and 2, 32 anchor items were selected: 16 overlapping items from both kindergarten and Grade 2 for Grade 1 and 16 overlapping items from both Grades 1 and 3 for Grade 2. The easiest and most difficult anchor items from each grade level were then moved back to the unique item pool to establish the upper and lower bounds of item difficulty for each grade level, and all items in the unique item pool were ordered by item difficulty from easiest to most difficult. Next, the remaining anchor items for each grade level were ordered with respect to their item difficulties from easiest to most difficult. Third, the items were organized into 20 blocks of five items each, and blocks were ordered by their average difficulty from easiest to most difficult. Items were randomly sampled from each block, with sampling beginning from a different starting point for each block to maximize randomization for different forms; the selected items were then placed in alternate forms for each TOY. We provide indices of form equivalence for the BOY, MOY, and EOY alternate forms for FEP by grade level in Table 9. The FEP form difficulties were -0.16, -1.98, -2.29, and -2.53 for Grades K, 1, 2, and 3, respectively.

Table 9: FEP Evidence of Alternate Form Equivalence, by Grade Level and TOY

	BOY		MOY		EOY	
Grade	**Form Difficulty**	**Number of Syllables**	**Form Difficulty**	**Number of Syllables**	**Form Difficulty**	**Number of Syllables**
K	–0.16	204	–0.1	204	–0.16	204
1	–1.98	226	–1.98	226	–1.98	226
2	–2.29	272	–2.29	272	–2.29	272
3	–2.53	324	–2.53	324	–2.53	324

Passage Development

mCLASS Lectura contains two subtests with passages: Fluidez en la Lectura Oral (FLO; oral reading fluency) and ¿Cuál Palabra? (CP; Maze; comprehension).

To support passage development, native Spanish speakers from diverse backgrounds across multiple Spanish-speaking countries (e.g., Argentina, Chile, Cuba, El Salvador, México, Nicaragua, and U.S./Puerto Rico) were hired to write passages of authentic and culturally relevant Spanish text. In particular, passage writers were given guidelines about specific passage features to attend to during their writing of the passages, including (a) grade-level appropriate readability, as measured by a formal readability index (such as Flesch-Kincaid, Crawford (1985), or other readability scores that take into consideration syntactic and semantic complexity when evaluating passage difficulty); (b) Spanish Lexile scores, which provide another index of passage difficulty and grade-level appropriateness; and (c) sensitivity of content, such that passages should include issues of diversity in terms of socioeconomic status, disability status, race/ethnicity, family structure, etc. As part of the production process for the text passages, we also attended to font size to ensure that selected fonts not only had easily distinguishable letters (e.g., capital I from lowercase l) but also reflected research suggesting font size and line length may interfere with reading comprehension (Katzir et al., 2013); as a result of this research, font sizes for the connected text passages

get progressively smaller from Grade 2 to Grade 5.

The process used to develop the passages, as well as passage reading levels and empirical evidence supporting the equivalency of the passages by grade level, is provided next. CTT and IRT (as previously described) were conducted to examine item difficulty and discrimination of the passages.

Fluidez en la lectura oral (FLO; Reading Fluency)

Based on explicit passage development guidelines described previously, nine passages were written for each grade level that were then pilot-tested with a varying range of students (n = 30 to 100 students per passage) and descriptive statistics for the words read correctly and for accuracy (number of words read correctly/total number of words read x 100) scores were calculated. Two methods were used to equate the passages: (1) the Delta method, which was calculated by subtracting the grand mean of all passages from each specific passage mean (Christ & Ardoin, 2009), and (2) an IRT-based equating approach that focused more on passage accuracy than words read correct (Powell-Smith et al., 2010). To equate the passages, we consulted the plots from both of these methods and selected the three passages from each grade level that were the most similar. Passage length (i.e., EOY passages should be the longest), and text type (i.e., a mix of narrative and informational passages) also influenced our passage selection process. We present the IRT passage difficulties (and corresponding standard errors of the passage difficulties) for each grade level, by TOY, in Table 10.

Table 10: Readability estimates and passage difficulties based on IRT equating methods for FLO benchmark passages by grade level and TOY

	BOY			MOY			EOY		
Grade	**Lexile**	**Crawford Score**	**Passage Difficulty**	**Lexile**	**Crawford Score**	**Passage Difficulty**	**Lexile**	**Crawford Score**	**Passage Difficulty**
1	210L-400L	2.5	0.13 (-0.37)	210L-400L	1.7	0.07 (0.39)	10L-200L	1.4	-0.06 (0.41)
2	210L-400L	2.6	0.06 (0.39)	410L-600L	2.7	0.02 (0.40)	610L-800L	2.8	-0.05 (0.27)
3	410L-600L	3.3	0.11 (0.31)	410L-600L	3.6	-0.06 (0.32)	810L-1000L	3.6	-0.12 (0.42)
4	610L-800L	4.6	0.44 (0.36)	610L-800L	4.7	0.40 (0.47)	610L-800L	4.6	0.39 (0.56)
5	610L-800L	4.9	0.04 (0.63)	1010L-1200L	5.0	0.13 (0.55)	810L-1000L	5.1	-0.15 (0.38)
6	810L-1000L	5.9	0.01 (0.54)	1010L-1200L	5.6	0.04 (0.40)	1030L-1335L	5.5	-0.02 (0.55)

Note: Cells that are shaded gray indicate passage statistics for informational texts.

¿Cuál palabra? (CP; Fluency; Reading Comprehension)

Development of Spanish passages for this reading comprehension subtest followed the same structure as described previously for Fluidez en la lectura oral (FLO; Reading Fluency).

Six passages for Grades 1–6 were developed for potential inclusion in the benchmark pool and the results from three equating methods were compared to help identify roughly equivalent passages for use at BOY, MOY, and EOY for each grade level. The three equating methods compared included: (1) the Delta

method, in which the grand mean percentage of items correct (across all passages within a grade level) was subtracted from the mean percentage of items correct for each unique passage (Christ & Ardoin, 2009); (2) a Rasch-based IRT approach, in which passage difficulty was determined by calculating whether the percentage of items correct was greater than the overall median percentage of items correct; and (3) a Rasch testlet IRT approach, in which the item difficulties were calculated concurrently, taking into account the nested structure of the data (i.e., all of the items are nested within one passage). Based on a comparison of the plots from these three equating methods, we used the average of the item difficulties for each passage obtained using the Rasch testlet model to equate the passages, and we selected the three passages with the closest average item difficulties as the benchmark forms. We present those passage difficulties by grade level and TOY in Table 11. Based on preliminary data from the item calibration study in 2020–2021, we opted to make Grade 1 CP an optional subtest to provide educators with more information about students' Spanish reading comprehension skills without informing the Grade 1 composite scores.

Table 11: Readability statistics and passage difficulties based on IRT equating methods for ¿Cuál Palabra? (CP) Benchmark passages by grade level and TOY

	BOY			MOY			EOY		
Grade	**Lexile**	**Crawford Score**	**Passage Difficulty**	**Lexile**	**Crawford Score**	**Passage Difficulty**	**Lexile**	**Crawford Score**	**Passage Difficulty**
2	210L–400L	2.6	−0.35 (0.28)	210L–400L	2.9	−0.36 (0.27)	410L–600L	2.8	−0.37 (0.27)
3	410L–600L	3.7	−0.91 (0.36)	410L–600L	3.8	−0.93 (0.31)	410L–600L	3.7	−1.11 (0.33)
4	610L–800L	4.7	−0.74 (0.35)	810L–1000L	4.6	−0.71 (0.32)	610L–800L	4.6	−0.77 (0.32)
5	1010L–1200L	5.6	−0.75 (0.32)	1010L–1200L	5.8	−0.71 (0.32)	1010L–1200L	5.7	−0.63 (0.33)
6	1210L–1400L	6.9	−0.90 (0.40)	1010L–1200L	5.9	−1.01 (0.45)	1210L–1400L	6.3	−0.80 (0.39)

Results

Descriptive Statistics

Table 12 displays the descriptive statistics for mCLASS Lectura subtest scores by grade and TOY from the 2021–2022 mCLASS Lectura development study. Overall, the results indicate that in each grade level, average scores on subtests increased over time, and the standard deviation decreased on all subtests except for FLO_ACC. The decrease in FLO_ACC suggests that more students were able to read the words correctly over time, which is expected because the benchmark forms were designed to be of equal difficulty. There were, of course, some exceptions to the trend of the average score: FSS in Kindergarten, in which the average score remained stable over time, FLO_WRC in Grade 4, and CP in Grade 3. This could be

due to the fluctuation in the samples at different TOYs, passage (or form) effects, or both.

Table 12: Sample A: Descriptive Statistics of mCLASS Lectura Subtest Scores by Grade and TOY

		BOY			MOY			EOY		
Grade	**Measure**	**N**	**Mean**	**SD**	**N**	**Mean**	**SD**	**N**	**Mean**	**SD**
K	FNL	4084	11.50	11.81	5640	23.61	13.84	5666	31.92	15.24
	FSS	4266	22.19	15.66	6136	35.11	15.95	6328	45.98	16.59
	FSL	4454	11	10.28	6586	22.69	12.42	6596	30.72	14.28
	LSS	3353	3.36	6.39	5096	11.36	10.78	5327	20.45	13.73
	FEP	4159	2.45	5.51	6235	7.95	9.75	6322	16.02	14.72
1	FNL	4621	26.60	13.75	5006	34	14.29	5051	39.54	15.05
	FSS	4754	31.76	14.43	6308	42.46	15.47	6061	50.05	16.48
	FSL	5094	25.32	12.73	6361	34.38	13.51	6179	40.91	14.90
	LSS	4356	16.75	13.87	5166	26.16	15.54	5164	34.85	16.56
	FEP	4816	12.21	13.23	6338	21.09	16.74	6154	29.84	19.64
	FLO_WRC	4698	14.04	15.77	6212	21.90	19.40	6085	33.33	25.75
	FLO_ACC	4694	54.30	33.65	6212	70.51	29.91	6085	77.79	27.62
2	FEP	5084	19.83	16.23	6182	26.6	18.42	5895	34.1	20.76
	FLO_WRC	5081	35.32	25.91	6192	53.08	30.24	5904	57.43	31.77
	FLO_ACC	5069	77.39	28.01	6192	87.77	21.12	5904	89.77	18.52
	CP	2130	2.71	3.49	4627	3.62	3.66	4369	4.33	4.52
3	FEP	2725	22.55	13.52	3534	28.67	15.26	3083	33.49	16.18
	FLO_WRC	2705	44.49	24.99	3582	60.89	29.1	3170	83.37	32.3
	FLO_ACC	2704	86.33	17.87	3582	91.28	16.17	3170	95.35	10.7
	CP	924	5.58	4.61	2598	4.22	4.38	2285	6.6	5.54
4	FLO_WRC	1417	60.2	27.76	1878	67	27	1517	73.87	26.64
	FLO_ACC	1417	90.21	14.74	1878	93.73	11.56	1517	94.44	9.68
	CP	475	5.66	4.26	1532	6.22	5.06	1223	7.8	5.89

		BOY			MOY			EOY		
Grade	**Measure**	**N**	**Mean**	**SD**	**N**	**Mean**	**SD**	**N**	**Mean**	**SD**
5	FLO_WRC	1116	73.61	23.52	1463	95.61	31.6	1102	95.62	35.7
	FLO_ACC	1116	93.59	11.03	1463	96.67	7.20	1102	95.99	7.46
	CP	509	5.58	4	1229	8.25	5.96	942	8.77	6.42

Correlations

The correlations between mCLASS Lectura subtests by grade level and TOY from the 2021–2022 study are summarized as follows. We provide empirical evidence that the skills assessed by mCLASS Lectura subtests are reasonably correlated to provide educators with an overall estimate of students' Spanish literacy skills but are not so highly correlated to yield redundant information about students' developing Spanish literacy skills. See Tables 13–15 for correlations among mCLASS Lectura subtests by grade level and TOY.

Examination of the data in Table 13 for kindergarten, for example, reveals strong correlations between subtests of letter naming fluency and letter sound fluency (FNL and FSL, respectively) that are consistent with prior research (Anthony et al., 2006), as well as moderate to strong (r = .61–.76) correlations between LSS (syllable reading), FNL, and FSL, which are to be expected theoretically because knowledge of letter sounds are necessary to decode Spanish syllables. Moreover, the increase in the magnitude of the correlations among these three subtests by TOY is to be expected as well, given that by EOY in Kindergarten students should have received instruction on all Spanish letter names and letter sounds and practiced blending sounds into larger units. Similar trends are observed between these three subtests and FEP (word reading), with particularly strong correlations observed between LSS (syllable sounds) and FEP across all three timepoints. These strong correlations are also expected because the real words on FEP comprise many of the Spanish syllables that appear in LSS. Also consistent with the literature (Miguez-Álvarez et al., 2021) are the smaller correlations between FSS and all other subtests; however, FSS is included within mCLASS Lectura as a necessary subtest of phonological awareness that requires students to engage in response processes similar to other English and Spanish literacy CBMs (Alonzo et al., 2013; Imagination Station, 2016). Collectively, these results are as expected based on prior research and suggest stronger empirical relationships among some subtests (e.g., subtests measuring varying complexity of the same construct, such as alphabetic understanding), low to moderate correlations among other pairs of subtests, and, of equal importance, no correlations of large magnitude (i.e., < .90) that might suggest multicollinearity and redundancy in the measurement of students' Spanish literacy skills.

Table 13: Correlations among mCLASS Lectura Subtest Scores for Kindergarten Students

TOY	Subtest	FSS	FSL	LSS	FEP
BOY	FNL	.32	.77	.61	.51
	FSS		.36	.24	.15
	FSL			.65	.51
	LSS				.82
MOY	FNL	.33	.78	.71	.62
	FSS		.40	.31	.25
	FSL			.73	.62
	LSS				.90
EOY	FNL	.37	.78	.73	.67
	FSS		.46	.37	.33
	FSL			.76	.68
	LSS				.92

Relationships of similar magnitude were observed among mCLASS Lectura for Grade 1 (see Table 14). As in kindergarten, results from Grade 1 reveal moderate correlations among FNL, FSL, and LSS (r = .66–.74). Stronger correlations were observed between LSS and FEP for Grade 1 (r = .90–.92); however, we would argue that both subtests yield instructionally useful and different information, as it is possible that students may be able to decode single syllables fluently but may struggle decoding multisyllabic words comprised of three, four, or five or more syllables. Similarly, although the correlations between FEP and FLO_WRC (words read correctly in the context of connected text passages) are large (r = .92–.94), these subtests provide instructionally valuable information about a student's ability to decode words in isolation as well as in the context of connected text. Research (Jenkins et al., 2003) has shown, for example, that some readers, particularly struggling readers, may be able to use the context of passages to support their word reading accuracy.

Table 14: Correlations among mCLASS Lectura Subtest Scores for Grade 1 Students

TOY	Subtest	FSS	FSL	LSS	FEP	FLO_ WRC	FLO_ ACC
BOY	FNL	.31	.70	.68	.60	.58	.65
	FSS		.34	.29	.26	.24	.30
	FSL			.67	.57	.54	.64
	LSS				.91	.88	.82
	FEP					.93	.77
	FLO_WRC						.74
MOY	FNL	.31	.70	.68	.63	.61	.61
	FSS		.31	.29	.26	.22	.26
	FSL			.66	.57	.53	.59
	LSS				.91	.86	.78
	FEP					.92	.72
	FLO_WRC						.68
EOY	FNL	.34	.70	.65	.62	.61	0.59
	FSS		.38	.34	.31	.26	.31
	FSL			.65	.56	.55	.58
	LSS				.90	.86	.81
	FEP					.92	.73
	FLO_WRC						.69

Similar trends can be seen among fluency rate (FLO_WRC), reading accuracy (FLO_ACC), and reading comprehension scores across TOYs in Grades 2–5 (see Table 15). Similar to Grade 1, for example, strong correlations were observed between FEP and FLO_WRC in Grades 2 and 3 (r = .89–.94), but again these subtests yield instructionally different information for teachers about student's facility with word reading in different contexts. Correlations between FEP and FLO_ACC, and FLO_WRC and FLO_ACC, for Grades 2 and 3 were also moderate, ranging from r = .54–.66 and r = .59–.71, respectively, suggesting that decoding and accuracy are not inextricably linked and that it may be important for educators to include both constructs

as instructional goals. Correlations between FLO_WRC and FLO_ACC were slightly lower for Grades 4–5, ranging from r = .52–.74, which may be partly due to small sample sizes in the upper grades. Finally, correlations between measures of word reading and reading comprehension (FEP and CP) and fluency and reading comprehension (FLO_WRC and CP) were moderate (r = .53–.61 and r = .54–.66), while correlations between reading accuracy and reading comprehension (FLO_ACC and CP) were noticeably lower, ranging from r = .22–.38. This finding, although potentially surprising given the need to accurately decode text to support reading comprehension, is commensurate with prior Spanish literacy research (López-Escribano et al., 2013) in which it has been argued that because of its transparent orthography, readers of Spanish may be able to decode text far beyond a level at which they can comprehend it.

Table 15: Correlations among mCLASS Lectura Subtest Scores for Grade 2–5 Students

TOY	Grade	Subtest	FLO_WRC	FLO_ACC	CP
BOY	2	FEP	.91	.64	.54
		FLO_WRC		.71	.54
		FLO_ACC			.31
	3	FEP	.89	.61	.61
		FLO_WRC		.66	.66
		FLO_ACC			.38
	4	FLO_WRC	–	.67	.52
		FLO_ACC		–	.27
	5	FLO_WRC	-	.67	.50
		FLO_ACC		–	.20

TOY	Grade	Subtest	FLO_WRC	FLO_ACC	CP
MOY	2	FEP	.93	.60	.56
		FLO_WRC		.68	.56
		FLO_ACC			.28
	3	FEP	.91	.62	.54
		FLO_WRC		.66	.57
		FLO_ACC			.26
	4	FLO_WRC	–	.65	.62
		FLO_ACC		–	.35
	5	FLO_WRC	–	.54	.64
		FLO_ACC		–	.29
EOY	2	FEP	.94	.61	.57
		FLO_WRC		.63	.59
		FLO_ACC			.28
	3	FEP	.91	.54	.55
		FLO_WRC		.59	.58
		FLO_ACC			.25
	4	FLO_WRC	–	.62	.65
		FLO_ACC		–	.36
	5	FLO_WRC	–	.58	.68
		FLO_ACC		–	.34

Composite Score Development

Factor analytic methods were used to develop the mCLASS Lectura Composite Score to provide evidence that the dimensionality of the construct matches the dimensionality of the assessment data.

Determining Subtest Weights for Computing Composite Scores

The mCLASS Lectura Composite Score is a linear combination of scores on mCLASS Lectura subtests that provides an estimate of overall student Spanish literacy skills. To compute composite scores for mCLASS Lectura, we used a Confirmatory Factor Analysis (CFA) approach. For each grade, we used a balanced statistical modeling approach, considering not only the empirical model fit but also theories of Spanish literacy development and literacy assessment when building the models. The models were built iteratively, starting with a base model for each grade, where all mCLASS Lectura subtests for that grade were loaded on the common reading factor. See Table 16 for a summary of subtests by grade. Thismodel was extended by modeling different types of covariances to account for the theoretical or measurement-driven relationships among the scores.

Table 16: mCLASS Lectura Subtests Available by Grade

Grade	FNL	FSS	FSL	LSS	FEP	FLO	CP
K	X	X	X	X	X		
1	X	X	X	X	X	X	
2					X	X	X
3					X	X	X
4						X	X
5						X	X
6*						X	X

**For Grade 6, additional study is planned to expand the sample size and build on the preliminary evidence of reliability, validity, and classification accuracy analyzed during the 2021-2022 school year.*

We hypothesized that a single dimension of Spanish reading skill is measured by the mCLASS Lectura subtests, so we fit a one-factor CFA model to the data for each grade at each TOY. We also expected to see higher correlations among some subtests because the foundational reading skills they measured are more similar than those by other subtests, or because the scores were derived from the same subtests. To account for these effects, we correlated the residuals (i.e., modeled covariances) of some subtests. For kindergarten, we correlated residuals of LSS and FEP because both subtests involve blending sounds into larger units (syllables and words, respectively). For Grade 1, we correlated the residuals of FEP and FLO_WRC because they both measure word reading skills, although the former asks students to read isolated real words out of context and the latter asks students to read real words in connected text. We also correlated the residuals for FNL and FSL because these are both assessments of letter knowledge. For Grade 2, we correlated FLO_WRC and FLO_ACC because their scores were derived from the same subtest. For Grade 3, we correlated FLO_ACC and CP to account for the over-estimation of their correlation by the base model. For Grades 4 and 5, as there were only three measures, we did not make any residuals correlated.

The final model for each grade level was determined by comparing model fits. Model fit was evaluated using the CFI (Bentler, 1990; acceptable fit ≥ .95), root mean square error of approximation (RMSEA; Browne

& Cudeck, 1993; acceptable fit ≤ .06), the standardized root mean square residual (RMSR; Hu & Bentler, 1998; acceptable fit ≤ .10), Akaike information criterion (AIC; Akaike, 1974; lower values, relative to other nested models, are better), and Bayesian information criterion (BIC; Schwarz, 1978; lower values, relative to other nested models, are better). Models were fit to data collected in the fall of 2021, using maximum likelihood estimation.

The results of the CFA models, by grade level and TOY, are presented in Table 17.

The model fit statistics indicate that most of the models had an excellent fit. Some models had slightly higher RMSEA, which could be due to a small degree of freedom (Kenny et al., 2015) and/or sample fluctuations.

Table 17: Model Fit Statistics for mCLASS Lectura Confirmatory Factor Analysis

Grade	Specification	TOY	N	CFI	RMSEA	SRMR
K	All mCLASS Lectura subtests + LSS~FEP	BOY	2756	.996	.031	.017
		MOY	4252	.997	.050	.013
		EOY	4403	.995	.070	.014
1	All mCLASS Lectura subtests + FEP~FLO_WRC; FNL~FSL	BOY	3641	.977	.103	.095
		MOY	4335	.980	.098	.034
		EOY	4574	.985	.087	.034
2	All mCLASS Lectura subtests + FLO_WRC~ FLO_ACC; FLO_ACC~CP	BOY	1434	>.999	.000	.000
		MOY	4335	>.999	.000	.000
		EOY	4030	>.999	.000	.000
3	All mCLASS Lectura subtests + FLO_ACC~CP	BOY	774	.988	.175	.011
		MOY	2467	>.999	.000	.025
		EOY	2065	>.999	.018	.003
4	All mCLASS Lectura subtests + FLO_WRC	BOY	410	>.999	.000	.000
		MOY	1426	.994	.07	.012
		EOY	1118	.999	.028	.007

Grade	Specification	TOY	N	CFI	RMSEA	SRMR
5	All mCLASS Lectura subtests + FLO_WRC	BOY	413	>.999	.000	.000
		MOY	1163	.997	.042	.009
		EOY	821	.993	.073	.023

We obtained the standardized factor loadings from the resulting best-fitting model and multiplied each of the factor loadings by the standard deviation of the corresponding subtest score. The product of this calculation is the weight for each mCLASS Lectura subtest that indicates its relative contribution to the mCLASS Lectura Composite Score. The composite score is a scaled linear combination of the weighted subtest scores.

Developing Cut Scores

mCLASS Lectura data collected over the course of the 2021–2022 school year were used to establish score ranges that correspond to performance levels (e.g., *Above Benchmark, Benchmark*, *Below Benchmark*, and *Well Below Benchmark*) for the mCLASS Lectura Composite Score and each subtest. Analysis was conducted to ensure that data from mCLASS Lectura yield consistent and trustworthy inferences about student placement into a performance level based on their demonstration of early Spanish literacy skills, so that they receive the necessary level of instructional supports.

To identify at-risk students, cut scores were generated from Sample C data for each mCLASS Lectura subtest and for composite scores at each TOY. mCLASS Lectura cut scores were determined using WM-AP in kindergarten, SELSp in Grades 1–3, and SRSp in Grades 4–5 as the external criterion measures. The first score, the risk cut score, classifies students who are well below benchmark in their performance and at risk for reading difficulties, including dyslexia.

The cut scores were calculated using Receiver Operating Characteristic (ROC) curve analyses, which describe the relation between true positive rates (i.e., scores that correctly identify students who are not on track for attaining proficiency) and false positive rates (i.e., scores that falsely indicate that a student was not on track for attaining proficiency). In this case, the ROC results characterize the extent to which mCLASS Lectura scores accurately predicted performance on the external Spanish criterion measures (i.e., WM-AP for kindergarten and Star Spanish for Grades 1–5). ROC analyses yield an area under the curve (AUC) estimate that summarizes the classification accuracy for the screening test of interest (i.e., mCLASS Lectura). An AUC of .5 indicates that the test predicts performance on the external criterion measure no better than chance, whereas an AUC of 1.0 indicates that a test is perfectly predictive (Habibzadeh et al., 2016).

In addition to the AUC, ROC analyses provide information about the sensitivity and specificity of a screener. The sensitivity index represents a proportion (ranging from 0 to 1) of the total number of students who were truly at risk on the criterion who were identified by the screener as being at risk. Specificity, also represented as a proportion, represents the proportion of truly healthy readers who are accurately identified as not at risk by the screener (i.e., identified as okay). Sensitivity can also be interpreted as the probability (likelihood) that a student who meets the criterion goal has been identified as such by the screener.

Although sensitivity and specificity are stable indicators of screening effectiveness regardless of the prevalence of reading difficulties in the population (Pepe, 2003), an important determinant of sensitivity and specificity that does not affect the AUC is how the cut score for the screener is set. mCLASS Lectura cut scores that balance sensitivity and specificity (to the greatest extent possible) have been selected, given their complementary role in a prevention model in education. Specifically, balancing both statistics results in maximizing the proportion of students correctly identified for intervention without under-identifying students correctly identified as not in need of intervention. Thus, wherever possible, recommended cut scores for the mCLASS Lectura Composite Score and for each of the mCLASS Lectura subtests were set to maximize sensitivity while maintaining specificity at or above .80. More specifically, for each benchmark, the cut was set at the score with the highest sensitivity among scores with specificity at or above .80. When there was a large difference between sensitivity and specificity, or when no specificity met the threshold of .80 or greater, we looked for the cut that yielded the highest combination of sensitivity and specificity to balance the goals of providing intervention to students who need it and keeping instructional demands on teachers reasonable.

Regardless of criterion measure, the 20th percentile rank cut is intended for use in identifying students who are well below benchmark, or at risk for not meeting EOY learning goals, and in need of intensive intervention. Students falling below this cut may also be at risk for reading disabilities, including dyslexia. The 40th percentile cut is intended for use in identifying students who are below benchmark, at some risk of not meeting EOY learning goals, and in need of some support. The sensitivity, specificity, and AUC values for the at-risk and some-risk cut scores for mCLASS Lectura are included in the Appendix, by grade level and TOY.

Chapter 2: Reliability of mCLASS Lectura

Reliability is generally described as the consistency with which an assessment measures the target skill(s) of interest; reliability statistics present information about the precision of an instrument, often expressed as a ratio. A test with perfect score precision has a reliability coefficient equal to 1, meaning that 100% of the variation among students' scores is attributable to variation in the trait or skill the test measures, and none of the variation is attributable to error. Perfect reliability is unattainable in educational measurement; a test with a reliability coefficient of 0.90 is more likely. On such a test, 90% of the variation among students' scores is attributable to the trait or skill being measured, and 10% is attributable to errors of measurement. If the trait or skill were measured a second time, students' scores would fluctuate to some degree; that is, scores on the second test would not be perfectly consistent with the same students' initial scores, but the scores would be close enough that educators could have confidence that the assessment was measuring the target skill(s) consistently. Reliability is an essential characteristic of screening and progress monitoring assessments that are used for instructional decision-making; if results are spurious and unreliable, inappropriate decisions might be made.

This section provides details on several types of reliability evidence for the mCLASS Lectura Composite Score and subtests:

- Internal consistency reliability refers to the degree of confidence in the precision of scores from a single measurement (i.e., the extent to which scores on the items are related to the students' overall score on the assessment).
- Standard error of measurement (SEM) examines the extent to which the mCLASS Lectura Composite Score or subtest score is likely to fluctuate due to chance or irrelevant factors.
- Inter-rater reliability refers to the degree to which different raters consistently score student responses.

Internal Consistency Reliability

To establish evidence of test score reliability for mCLASS Lectura subtests, we examined the internal consistency of the composite score for each grade. Salvia, Ysseldyke, and Witmer's (2016) standards for reliability were used to evaluate the reliability data for mCLASS Lectura. According to these standards, a minimum reliability of 0.60 is ideal for making educational decisions about groups of students, a minimum of 0.70 suggests adequate reliability generally, a minimum of 0.80 is ideal for screening decisions, and a minimum of 0.90 is required for important educational decisions concerning an individual student. Coefficient α (Cronbach, 1951) is reported for internal consistency. Table 18 shows the sample size, the values of coefficient alpha, and its 95% confidence interval of the composite score for all grades at each TOY.

Overall, internal consistency reliability ranged from α = .75–.90 across Grades K–3, which is characterized as moderate to strong reliability (Salvia et al., 2016). Internal consistency reliability was highest for Grade 1 at BOY, MOY, and EOY, which was expected because students in Grade 1 were administered the most subtests, and the composite score is based on a weighted combination of the seven subtests. These internal consistency reliability estimates meet the National Center on Intensive Intervention (NCII) academic screening technical standard for reliability for which the lower bound of the 95% confidence interval around the reliability estimate met or exceeded .70 (NCII, 2018). Internal consistency reliability estimates in Grades 4 and 5 are lower (α= .48–.65), which is not surprising given that there were fewer

subtests contributing to the composite score. On average, internal consistency reliability estimates were greatest at BOY compared to MOY and EOY, and greater for Grade 4 than Grade 5; these results may be attributable, in part, to the increasingly weaker correlations among subtests and the smaller sample sizes in the upper-grade levels.

Table 18: Internal Consistency of mCLASS Lectura Composite Score for Grades K-5 at BOY, MOY, and EOY

		BOY			MOY			EOY		
Grade	**N of Subtests**	**Sample Size**	**α**	**95% CI**	**Sample Size**	**α**	**95% CI**	**Sample Size**	**α**	**95% CI**
K	5	2766	.76	.74, .77	4262	.84	.83, .85	4413	.88	.87,.89
1	7	3651	.88	.87, .88	4345	.89	.89, .90	4584	.90	.90,.91
2	4	1444	.81	.80, .81	4345	.81	.80, .81	4040	.81	.80,.81
3	4	784	.81	.80, .83	2477	.79	.79, .80	2075	.75	.74,.76
4	3	420	.63	.60, .66	1436	.62	.60, .64	1128	.60	.57, .62
5	3	423	.65	.61, .68	1173	.48	.45, .51	831	.49	.46, .52

In sum, these results demonstrate that mCLASS Lectura is a reliable assessment for making educational decisions.

Standard Error of Measurement

In addition to internal consistency reliability, we also computed the SEM for the mCLASS Lectura Composite Score and all subtests to provide additional evidence for score precision. We computed the SEM using the formula $\sigma_x * \sqrt{1 - \rho xx}$, where σ_x is the standard deviation of the observed score and ρxx is its reliability coefficient. It can be seen that the SEM is influenced by both the standard deviation of the score and its reliability. For the composite score, we used the coefficient α to represent its reliability; for the subtests, we used their inter-rater reliability. Note that the standard deviation of the composite score is constant (SD = 40) across grades and TOYs, which means that any observed differences in their SEMs across grades and TOY were due to differences in reliability. The SEM for the mCLASS Lectura Composite Score for each grade level and TOY is presented in Table 19. The composite score SEMs are relatively stable in Grades 1 to 3, ranging from 12 to 13 in Grade 1, around 17 in Grade 2, and from 17 to 20 in Grade 3. The composite score SEMs are slightly larger in Grades 4 and 5, which was caused by their relative lower reliabilities and affected by the smaller number of subtests in these grade levels. SEMs vary the most in kindergarten with a range of 13–19 and decrease over time; this is likely a result of students having a wide range of skills as incoming kindergarten students at BOY that is moderated by the instruction they received over the course of the school year.

Table 19: Standard Error of Measurement (SEM) for mCLASS Lectura Composite Score by Grade and Time of Year

Grade	BOY	MOY	EOY
K	19.59	16.00	13.86
1	13.86	13.27	12.65
2	17.42	17.43	17.44
3	17.44	18.34	20.00
4	24.34	24.97	25.30
5	23.67	28.84	28.58

The SEMs for the mCLASS Lectura subtest scores are influenced by both the reliability of that subtest score and its standard deviation. Table 20 shows the SEMs of five subtests on which we obtained their inter-rater reliability. For almost all of the subtests, the SEMs tend to increase from kindergarten to Grade 2, which is partly due to the increasing standard deviations for the mCLASS Lectura subtests across the grades.

Table 20: Standard Error of Measurement (SEM) for mCLASS Lectura Subtest Scores by Grade

Subtest	K	1	2	3	4	5
FNL	1.45	3.20	–	–	–	–
FSS	4.07	3.21	–	–	–	–
FSL	1.47	1.91	–	–	–	–
LSS	2.18	3.87	–	–	–	–
FEP	1.15	1.06	2.02	1.91	–	–
FLO_WRC	–	2.94	6.70	3.39	3.95	4.24

Inter-rater Reliability

Inter-rater reliability was estimated for all mCLASS Lectura subtests except CP and the composite score. Although mCLASS Lectura is a low inference measure, some human judgment is required in order to produce student scores. For example, assessors must decide whether pronunciation of a sound or word is correct or incorrect. Therefore, it is useful to consider evidence of inter-rater reliability, or evidence that two independent raters score responses from the same student consistently. To estimate inter-rater reliability, we calculated intraclass coefficients (ICCs) for a subsample of students. Specifically, we conducted a one-way random-effects analysis and used consistency as the criterion. The one-way random-effects model was used because students were rated by different sets of raters. We focused on consistency because scores on mCLASS Lectura subtests can range from 0 to 245, which makes it hard for raters to reach absolute agreement. These ICCs, in other words, provide empirical evidence of "how observers or judges record and evaluate [student performance] data" (American Educational Research Association

(AERA) et al., 2014, p. 16).

Inter-rater reliability analysis was conducted at BOY and MOY during the 2021–2022 school year. One data collector administered each subtest while the other observed and recorded responses simultaneously on a separate device. The total correct raw scores recorded by the two data collectors were correlated to calculate an index of inter-rater reliability. According to Koo and Li (2016), a value of ICC below 0.50 indicates poor reliability, between 0.50 and 0.75 indicates moderate reliability, between 0.75 and 0.90 indicates good reliability, and any value above 0.90 indicates excellent reliability. A reliability coefficient of 0.99 demonstrates a high degree of inter-rater reliability for mCLASS Lectura subtests. Inter-rater reliability ranged from 0.94 to 1.0, indicating strong agreement between raters. Intra-class correlation coefficients are reported as an index for inter-rater reliability in Table 21, by grade and subtest. CP does not have an index of inter-rater reliability because it is administered online with multiple-choice items.

Table 21: Inter-rater Reliability of mCLASS Lectura Subtests by Grade

	K			1			2			3			4			5		
Subtest	**N**	**ICC**	**95% CI**	**N**	**ICC**	**95% CI**	**N**	**ICC**	**95% CI**	**N**	**ICC**	**95% CI**	**N**	**ICC**	**95% CI**	**N**	**ICC**	**95% CI**
FNL	40	.99	.98, .99	41	.95	.91, .97												
FSS	39	.94	.88, .97	39	.96	.92, .98												
FSL	40	.99	.97, .99	38	.98	.96, .99												
LSS	44	.96	.93, .98	49	.94	.89, .96												
FEP	41	.99	.97, .99	50	1.0	.99, 1.0	33	.99	.98, .99	30	.99	.97, .99						
FLO				42	.98	.96, .99	22	.95	.89, .98	50	.99	.98, .99	60	.98	.96, .99	27	.98	.96, .99

Summary

Taken together, the reliability evidence for mCLASS Lectura across grade levels is strong. Research into the reliability of mCLASS Lectura scores and scoring of the mCLASS Lectura subtests is ongoing, and regular addendums to this manual will continue to build the reliability argument for mCLASS Lectura.

Chapter 3: Validity

The validity of a test is the degree to which it assesses what it claims to measure. Formally, validity is defined as the degree to which evidence and theory support the interpretation of test scores entailed by proposed uses of tests (AERA et al., 1999). In other words, validity represents how confident we are that interpretations of test scores accurately represent what we believe they do (e.g., high scores on a comprehension assessment actually represent high comprehension skill). In this sense, validity is a way to describe a test's accuracy or utility.

Validity is not proven; evidence is collected to strengthen the assertion that a test accurately measures the desired construct(s). Validity was traditionally considered a property assessments themselves possessed; it was categorized as content, construct, and criterion validity. The current view, however, considers a more unified treatment under which validity evidence is collected to support test score interpretations for their intended or unintended use (Kane, 2001; Messick, 1989) and may be captured under a more general heading of evidence for construct validity. Determining the validity of a test involves the use of data and other information, both internal and external to the test instrument itself.

Criterion-related validity is the extent to which student performance on the assessment procedure being validated can estimate student performance on a criterion measure (AERA et al., 2014; Salvia et al., 2013) and includes both concurrent- and predictive-related validity evidence. Conceptualized broadly, criterion-related validity evidence for an assessment refers to the degree to which current outcomes are associated with outcomes on an external, conceptually related assessment; whether the evidence gathered is concurrent or predictive depends on when the external criterion assessments were administered (in relation to the target assessment). Evidence of concurrent validity is gathered when the target assessment and the external assessment are administered at approximately the same time, whereas evidence of predictive validity is gathered when performance on the target assessment is examined relative to performance when the external assessment is administered at some point in the future. Concurrent validity for mCLASS Lectura was evaluated by correlating mCLASS Lectura subtests and composite scores with scores from an external criterion assessment (i.e., WM-AP in kindergarten, SELSp in Grades 1–3, and SRSp in Grades 4–5) when both were administered at BOY, MOY, and EOY. Correlations for mCLASS Lectura subtests and composite scores with each criterion assessment by grade level are reported in Tables 22–27.

Predictive validity can also be seen as a means of validating that the intended construct has been captured; in addition, it serves as a means of validating the use of a measure for predicting performance at a later period (e.g., often the end of a grade). Predictive validity traditionally includes correlations, but when intended uses of a measure include identification of subgroups of students, then an evaluation of screening accuracy provides critical evidence that a measure is functioning as intended (Jenkins et al., 2007). Consequently, predictive validity for mCLASS Lectura was evaluated by (a) correlating mCLASS Lectura subtest and composite scores at BOY or MOY with WM-AP at EOY in Kindergarten, SELSp at EOY in Grades 1–3, and SRSp at EOY in Grades 4–5, all of which served as external criterion measures at BOY, MOY, and EOY; and (b) using signal detection methods and ROC curve analysis to examine the extent to which performance on mCLASS Lectura at BOY, MOY, and EOY accurately differentiates between students who did and did not demonstrate proficiency on the external criterion measures. Predictive correlations for mCLASS Lectura Composite Scores and subtests with each criterion measure by grade level are reported in Tables 28–33.

Concurrent Validity

Concurrent validity correlations for the kindergarten mCLASS Lectura Composite Score (CS) and subtests with WM-AP are presented in Table 22. Overall, the mCLASS Lectura Composite Score was moderately to strongly correlated with WM-AP at each TOY (i.e., r = .63 at BOY, .74 at MOY, and .73 at EOY). mCLASS Lectura subtest correlations ranged from small to strong with WM-AP, with the weakest correlations observed between FSS and WM-AP. Among the six subtests, LSS and FEP had the highest correlations with WM-AP, followed by FNL and FSL. These results are as we might expect given that alphabet knowledge, knowledge of letter-sound correspondences, and decoding require the same alphabetic understanding and decoding skills required to decode the orthographically regular nonsense words that constitute the WM-AP task.

Table 22: Concurrent Validity of mCLASS Lectura Composite Scores and Subtests with WM-AP in Kindergarten by TOY

	BOY			MOY			EOY		
Measure	**N**	**r**	**95% CI**	**N**	**r**	**95% CI**	**N**	**r**	**95% CI**
CS	226	.68	.53, .69	256	.74	.68, .79	297	.73	.67, .78
FNL	252	.56	.47, .64	276	.63	.55, .70	298	.63	.56, .70
FSS	252	.23	.11, .35	276	.27	.16, .38	299	.17	.06, .28
FSL	233	.57	.47, .65	260	.63	.55, .70	298	.63	.55, .69
LSS	235	.72	.65, .78	259	.80	.75, .84	299	.80	.76, .84
FEP	229	.67	.60, .74	257	.79	.74, .83	298	.78	.74, .82

The concurrent validity correlations for mCLASS Lectura and SELSp for Grade 1 at each TOY are presented in Table 23. The correlations between the mCLASS Lectura Composite Score and the SELSp scale score were moderate to strong, ranging from r = .68–.73. Among the nine mCLASS Lectura subtests, LSS, FEP, FLO_WRC, and FLO_ACC had the strongest correlations with the SELSp scale score at all TOYs, with correlations ranging from r = .60–.72. FNL and FSL were moderately correlated with the SELSp scale score (r = .48–.59) and FSS was weakly correlated (r = .19–25). There are several reasons why this might be the case. Whereas mCLASS Lectura is comprised of fixed-form subtests in which students are asked to engage in the same behavior for one full minute, SELSp is computer-adaptive, meaning that students may be asked to respond to multiple, different item types assessing multiple, different reading constructs within a given testing session; although comparison of the test blueprints suggests that the two assessments are measuring many of the same Spanish literacy constructs, the use of different—and, in the case of SELSp, more varying—item types may have contributed to the modest correlations observed. Further, mCLASS Lectura is a teacher-administered, one-on-one assessment where students are asked to produce sounds and read words aloud, as opposed to the SELSp tasks in which students are asked to select a response from a given set of answer choices in a group-administered, computer-based assessment. These significant differences in administration and types of knowledge and skills elicited (e.g., expressive versus receptive skills) may have influenced the relationships between these measures.

Table 23: Concurrent Validity of mCLASS Lectura Composite Scores and Subtests with SELSp in Grade 1 by TOY

	BOY			MOY			EOY		
Measure	**N**	**r**	**95% CI**	**N**	**r**	**95% CI**	**N**	**r**	**95% CI**
CS	564	.68	.64, .72	692	.73	.70, .77	809	.69	.65, .72
FNL	614	.59	.53, .64	723	.56	.51, .61	839	.52	.47, .57
FSS	607	.24	.16, .31	704	.19	.12, .26	814	.25	.19, .32
FSL	613	.55	.49, .60	723	.51	.45, .56	834	.48	.43, .53
LSS	611	.65	.60, .69	715	.72	.68, .75	831	.71	.68, .74
FEP	590	.63	.58, .67	713	.70	.66, .74	824	.67	.63, .70
FLO_WRC	568	.60	.55, .65	708	.67	.62, .71	820	.62	.58, .66
FLO_ACC	568	.65	.60, .70	708	.65	.60, .69	820	.71	.67, .74

The concurrent validity correlations for mCLASS Lectura and the SELSp scale score for Grade 2 at each TOY are presented in Table 24. The correlation between the mCLASS Lectura Composite Score and SELSp for Grade 2 was also strong at each TOY, ranging from r = .66–.69. At each TOY, FEP and FLO_WRC had the strongest correlations with the SELSp scale score (r = .66–.71), followed by FLO_ACC (r = .63–.66). Correlations between CP and the SELSp scale score were low (r = .37.49), which may be explained by two factors: (1) the proportion of students obtaining a score of 0 on CP in Grade 2 was high (i.e., 32.3% at BOY, 24.2% at MOY, and 23.8% at EOY), and (2) SELSp targets foundational skills and therefore does not include a comprehension component (Renaissance Learning, 2021).

Table 24: Concurrent Validity of mCLASS Lectura Composite Scores and Subtests with SELSp in Grade 2 by TOY

	BOY			MOY			EOY		
Measure	**N**	**r**	**95% CI**	**N**	**r**	**95% CI**	**N**	**r**	**95% CI**
CS	362	.66	.59, .71	473	.69	.64, .74	584	.67	.62, .71
FEP	481	.67	.61, .71	530	.67	.62, .71	674	.66	.62, .70
FLO_WRC	459	.67	.62, .72	529	.71	.66, .75	675	.68	.64, .72
FLO_ACC	450	.63	.57, .69	529	.65	.60, .70	675	.66	.62, .70
CP	385	.37	.29, .46	480	.49	.42, .56	620	.48	.42, .54

The concurrent validity correlations for mCLASS Lectura and the SELSp scale score for Grade 3 at each TOY are presented in Table 25. The correlation between the mCLASS Lectura Composite Score and the SELSp scale score for Grade 3 was moderately strong at each TOY, ranging from r = .64–.68. Among the four subtests, the correlations between FEP and SELSp, and FLO_WRC and SELSp, were the strongest and remained stable over time. The correlations between FLO_ACC and SELSp were moderate to strong at each TOY and the correlations between CP and SELSp were moderate to strong at BOY and EOY and moderate at MOY, yet stronger overall than in Grade 2.

Table 25: Concurrent Validity of mCLASS Lectura Composite Scores and Subtests with SELSp in Grade 3 by TOY

	BOY			MOY			EOY		
Measure	**N**	**r**	**95% CI**	**N**	**r**	**95% CI**	**N**	**r**	**95% CI**
CS	401	.64	.58, .70	370	.68	.62, .73	443	.68	.63, .73
FEP	431	.64	.58, .69	410	.63	.57, .69	448	.64	.58, .69
FLO_WRC	426	.63	.57, .68	410	.68	.62, .73	448	.68	.62, .73
FLO_ACC	425	.54	.47, .60	410	.63	.57, .68	448	.57	.50, .63
CP	406	.56	.49, .63	372	.48	.39, .55	472	.51	.44, .57

The concurrent validity correlations for mCLASS Lectura and the SRSp scale score for Grade 4 at each TOY are presented in Table 26. The correlations between the mCLASS Lectura Composite Score and the SRSp scale score for Grade 4 were moderate to strong at each TOY, ranging from r = .56–.64. The correlations between FLO_WRC and the SRSp scale score were also moderate-strong at each TOY, ranging from r = .57–.65. The correlations between CP and the SRSp scale score at MOY and EOY were moderate to strong as well (i.e., r = 63 at MOY and EOY). As expected, FLO_ACC had the weakest correlation with the SRSp scale score.

Table 26: Concurrent Validity of mCLASS Lectura Composite Scores and Subtests with SRSp in Grade 4 by TOY

	BOY			MOY			EOY		
Measure	**N**	**r**	**95% CI**	**N**	**r**	**95% CI**	**N**	**r**	**95% CI**
CS	376	.64	.57, .69	397	.65	.59, .70	331	.56	.48, .63
FLO_WRC	405	.65	.58, .70	412	.65	.59, .70	408	.57	.50, .63
FLO_ACC	405	.35	.27, .44	412	.34	.26, .43	408	.30	.21, .38
CP	376	.44	.35, .52	399	.63	.57, .69	362	.63	.57, .69

The concurrent validity correlations for mCLASS Lectura and the SRSp for Grade 5 at each TOY are presented in Table 27. The correlations between the mCLASS Lectura Composite Score and the SRSp scale score for Grade 5 were moderate to strong at each TOY, ranging from r = .54–.60. FLO_WRC and CP had much stronger correlations with the SRSp scale score than FLO_ACC, which could be explained by the non-normal distributions of FLO_ACC. Additional data will be collected for this grade level during the 2022–2023 school year.

Table 27: Concurrent Validity of mCLASS Lectura Composite Scores and Subtests with SRSp in Grade 5 by TOY

	BOY			MOY			EOY		
Measure	**N**	**r**	**95% CI**	**N**	**r**	**95% CI**	**N**	**r**	**95% CI**
CS	345	.54	.46, .61	391	.60	.54, .66	331	.60	.53, .67
FLO_WRC	368	.52	.44, .59	404	.60	.53, .66	338	.60	.53, .66
FLO_ACC	368	.22	.13, .32	404	.33	.24, .41	338	.36	.26, .45
CP	345	.49	.40, .56	391	.57	.50, .63	350	.64	.57, .70

In sum, moderate to strong positive relationships were observed with Star Spanish at each time point for all grade levels, with correlations ranging from r = .54–.74 for the mCLASS Lectura Composite Score. On average, correlations for the mCLASS Lectura subtests administered in Grades K, 2, 3, 4, and 5 (with a few notable exceptions) were moderate to strong. Correlations for the mCLASS Lectura subtests administered in Grade 1 were low to moderate, which (as described previously) may be explained by factors including (a) differences in the test design (e.g., fixed-form versus computer-adaptive) and/or (b) differences in the behaviors students needed to engage in to complete the tasks (e.g., expressive knowledge via production of responses versus receptive knowledge via selection of a response option from multiple-choice items). This hypothesis is also supported by the increase in the magnitude of concurrent correlations between mCLASS Lectura and Star Spanish across time, with the strongest concurrent correlations observed at EOY with the exception of FSL. Correlations for Grades 2 and 3 were moderate to moderately strong. Concurrent correlations between mCLASS Lectura and Star Spanish were strongest for the subtests measuring decoding skills of increasing complexity (e.g., syllable reading, word reading, and oral reading fluency). Finally, correlations for Grades 4 and 5 ranged from moderate to moderately strong with the strongest correlations for the mCLASS Lectura Composite Score and subtests measuring fluency and comprehension. Modest correlations were also observed between FLO_ACC and SRSp scale scores in Grades 4 and 5, which may be attributed to the non-normal distribution of FLO_ACC scores that resulted from the majority of the sample in these grades demonstrating high levels of accuracy while reading FLO connected text passages.

Overall, mCLASS Lectura Composite Scores and subtests had moderate to strong correlations with WM-AP and Star Spanish. Despite the differences in the subskills measured, item formats, and administration methods, the results suggest that mCLASS Lectura, WM-AP, SELSp, and SRSp measure roughly the same foundational Spanish literacy skills.

Predictive Validity

Predictive validity provides an estimate of the extent to which student performance on mCLASS Lectura predicts scores on external criterion assessments administered at a later point in time, operationally defined as more than 2 months after the initial administration of mCLASS Lectura. Estimated as the linear relationship between student performance on mCLASS Lectura and the criterion assessments, such predictive correlations are attenuated by time, because students gain skills in the interim between testing occasions, and also by differences in the content specifications of the assessments.

We present predictive validity of the mCLASS Lectura composite and subtest scores with the WM-AP for kindergarten in Table 28. Overall, correlations are greatest for the mCLASS Lectura Composite Score

(compared to the subtest scores) and for MOY (compared to BOY). Correlations between the mCLASS Lectura Composite Score and the WM-AP subtest were moderate to strong, ranging from r = .58.70. At the subtest level, correlations were largest for FNL and FSL at BOY (r = .50–.51) and largest for LSS and FEP at MOY (r = .70–.72). These correlations are as we might expect, with the majority of incoming kindergartners at BOY having knowledge of some letter names and sounds with decoding skills developing over the course of the school year as a result of instruction. Predictive correlations between LSS and FEP at BOY with WM-AP scores were noticeably lower (r = 0.40–0.44); and we hypothesize this may be due in part to the restricted range of scores on all of these measures at kindergarten BOY that likely resulted in attenuated correlations.

Table 28: Predictive Validity of mCLASSLectura BOY and MOY Composite Scores and Subtests with WM-AP in Kindergarten at EOY

	BOY			MOY		
Measure	**N**	**r**	**95% CI**	**N**	**r**	**95% CI**
CS	236	.58	.48, .66	243	.74	.67, .79
FNL	261	.51	.41, .59	266	.65	.57, .71
FSS	264	.34	.23, .44	266	.40	.30, .50
FSL	241	.50	.40, .59	247	.61	.53, .69
LSS	244	.44	.33, .53	246	.72	.66, .78
FEP	239	.40	.29, .05	244	.70	.63, .76

As shown in Table 29, the Grade 1 mCLASS Lectura Composite Score also had the highest predictive validity, compared to the subtest scores, with the SELSp scale score. Predictive correlations for the mCLASS Lectura Composite Score were moderate, ranging from r = .61–.67 at BOY and MOY, respectively. In contrast, the correlation coefficients for the mCLASS Lectura subtests at BOY ranged from r = .29.59 and at MOY ranged from r = .25–.66. Similar to kindergarten, mCLASS Lectura scores had stronger overall predictive validity at MOY than at BOY. Among the individual subtests at BOY, FLO_ACC predicted the SELSp EOY scale score most strongly, followed by LSS, FNL, FSL, FEP, and FLO_WRC; whereas at MOY, LSS was the strongest predictor, followed by FEP, FLO_ACC, and FLO_WRC. The predictive validity of FNL and FSL changed a little from BOY to MOY, but the two became weaker predictors than LSS, FEP, or FLO_WRC at MOY. FSS was the weakest predictor of the SELSp scale score for Grade 1 students.

Table 29: Predictive Validity of mCLASS Lectura BOY and MOY Composite Scores and Subtests with SELSp in Grade 1 at EOY

	BOY			MOY		
Measure	**N**	**r**	**95% CI**	**N**	**r**	**95% CI**
CS	590	.61	.55, .66	697	.67	.62, .71
FNL	636	.54	.49, .60	773	.54	.49, .59
FSS	630	.29	.22, .36	708	.25	.18, .32
FSL	633	.53	.47, .58	773	.50	.45, .55
LSS	631	.56	.50, .61	761	.66	.62, .70
FEP	611	.53	.47, .58	763	.63	.59, .67
FLO_WRC	591	.51	.44, .56	758	.60	.55, .64
FLO_ACC	591	.59	.54, .64	758	.62	.58, .67

In Grade 2, the mCLASS Lectura Composite Score had a moderate to strong correlation with the SELSp score at EOY (r = .58 and .66 at BOY and MOY, respectively; See Table 30). In contrast to kindergarten and Grade 1, several mCLASS Lectura subtests had higher predictive validity for the SELSp scale score than the mCLASS Lectura Composite Score, which could be explained by the differences in the sample sizes for the individual mCLASS Lectura subtests and the composite score. Among the individual subtests, FEP, FLO_WRC, and FLO_ACC were more predictive of the EOY SELSp scale score than CP at both time points. This is likely because the first three subtests are all measures of word reading skills, while CP is a measure of comprehension; there are no items on the SELSp designed to explicitly measure reading comprehension (Renaissance Learning, 2021).

Table 30: Predictive Validity of mCLASS Lectura BOY and MOY Composite Scores and Subtests with SELSp in Grade 2 at EOY

	BOY			MOY		
Measure	**N**	**r**	**95% CI**	**N**	**r**	**95% CI**
CS	398	.58	.51, .64	548	.66	.61, .71
FEP	562	.61	.55, .66	608	.63	.58, .68
FLO_WRC	567	.62	.57, .67	608	.68	.63, .72
FLO_ACC	558	.59	.54, .64	608	.65	.60, .69
CP	437	.34	.25, .42	565	.47	.40, .53

Among the individual measures in Grade 3, FLO_WRC was the strongest predictor of SELSp scale score at both BOY and MOY (see Table 31). At BOY, FEP was the next strongest predictor, followed by CP. FLO_ACC had the weakest predictive validity. At MOY, both FEP and FLO_ACC were the next strongest predictors, and CP had the lowest correlation with SELSp.

Table 31: Predictive Validity of mCLASS Lectura BOY and MOY Composite Scores and Subtests with SELSp in Grade 3 at EOY

	BOY			MOY		
Measure	**N**	**r**	**95% CI**	**N**	**r**	**95% CI**
CS	419	.56	.49, .62	412	.60	.54, .66
FEP	451	.53	.47, .60	452	.55	.48, .61
FLO_WRC	446	.54	.48, .61	452	.61	.55, .67
FLO_ACC	445	.43	.35, .51	452	.55	.48, .61
CP	427	.51	.43, .58	414	.47	.39, .54

Tables 32–33 show the predictive validity results of mCLASS Lectura for students in Grades 4–5 with SRSp at EOY. Students in these grades were administered only two subtests (i.e., FLO and CP). The predictive validity of the composite score was strong in Grade 4 and moderate in Grade 5. FLO_WRC and CP were the strongest predictors of SRSp scale score and FLO_ACC became the weakest predictor, which might be explained by the growing number of students who were able to achieve high scores on FLO_ACC, reducing the ability of this measure to differentiate among students with different levels of reading skills.

Table 32: Predictive Validity of mCLASSLectura BOY and MOY Composite Scores and Subtests with SRSp in Grade 4 at EOY

	BOY			MOY		
Measure	**N**	**r**	**95% CI**	**N**	**r**	**95% CI**
CS	368	.62	.55, .68	416	.65	.60, .71
FLO_WRC	401	.61	.55, .67	429	.65	.59, .70
FLO_ACC	401	.32	.23, .40	429	.33	.25, .42
CP	368	.42	.33, .50	444	.61	.54, .66

Table 33: Predictive Validity of mCLASS Lectura BOY and MOY Composite Scores and Subtests with SRSp in Grade 5 at EOY

	BOY			MOY		
Measure	**N**	**r**	**95% CI**	**N**	**r**	**95% CI**
CS	312	.53	.45, .61	370	.56	.49, .63
FLO_WRC	334	.52	.44, .60	379	.56	.49, .63
FLO_ACC	334	.25	.14, .34	379	.29	.19, .38
CP	312	.56	.48, .63	389	.58	.51, .64

In sum, similar to concurrent validity results, the strongest correlations were between SELSp and mCLASS Lectura subtests that measured decoding skills (i.e., LSS, FEP, and FLO_WRC). In Grades 1–3, the lowest subtest predictive correlations were observed between FSS and SELSp, again likely due to differences in how

phonological awareness skills are measured across the two assessments. In mCLASS Lectura, for example, students are asked to orally segment words into syllables, the scoring of which depends on human judgment and requires students to produce sounds out loud, whereas SELSp does not include any segmentation tasks but rather includes phonemic awareness tasks that focus on rhyming, blending, and the identification, isolation, and manipulation of phonemes in a multiple-choice format (Renaissance Learning, 2021).

Taken together, the results suggest that the mCLASS Lectura Composite Score and subtests can be used to accurately predict students' EOY performance on WM-AP, SELSp, and SRSp, established measures of overall reading skill.

Classification Accuracy

A common method for evaluating screening systems is accomplished by employing signal detection methods to evaluate how well a screening system (e.g., mCLASS Lectura) detects the occurrence of a later event or condition (e.g., demonstrating Spanish literacy proficiency on an established external criterion assessment, such as WM-AP, SELSp, and SRSp). Specifically, the goal of this methodology is to see how well mCLASS Lectura accurately differentiates between those who do or do not demonstrate proficiency on WM-AP in kindergarten, SELSp in Grades 1–3, and SRSp in Grades 4–5, where proficiency is defined as performing at or below the 20th percentile (at risk) and at or above the 40th percentile (some risk).

ROC analysis is commonly used to generate the four possible outcomes between the screening and criterion assessments:

- True Positive (Sensitivity): Of the students identified as at risk by WM-AP in kindergarten, SELSp in Grades 1–3, and SRSp in Grades 4–5, the proportion of students also identified as at risk by mCLASS Lectura.
- False Positive: Of the students identified as at risk by WM-AP in kindergarten, SELSp in Grades 1–3, and SRSp in Grades 4–5, the proportion of students identified as on track by mCLASS Lectura.
- True Negative (Specificity): Of the students identified as on track by WM-AP in kindergarten, SELSp in Grades 1–3, and SRSp in Grades 4–5, the proportion of students also identified as on track by mCLASS Lectura.
- False Negative: Of the students identified as on track by WM-AP in kindergarten, SELSp in Grades 1–3, and SRSp in Grades 4–5, the proportion of students identified as at risk by mCLASS Lectura.

The ROC curve has become the standard for evaluating the accuracy of screening systems using signal detection methods; and the AUC is the recommended index of accuracy (Smolkowski & Cummings, 2015). Specifically, the AUC represents the degree to which the screener (mCLASS Lectura) accurately differentiates students into the outcomes of interest on the criterion assessment (at risk or on track on WM-AP in kindergarten, on SELSp in Grades 1–3, and on SRSp in Grades 4–5).

In Kindergarten, the AUCs for the mCLASS Lectura Composite Score ranged from .80 to .96, with the majority falling at .85 or above. FNL, FSL, and FEP had higher AUCs than other subtests. LSS also had high AUCs except at BOY when predicting at-risk status on WM-AP at the 20th percentile. FSS had the lowest AUCs, which was not surprising given its low correlations with WM-AP among students in kindergarten. All of the sensitivity and specificity values for the composite score were above .70, and a few of them were above .90. The AUCs of the subtests showed that all subtests except FSS were strong predictors of SELSp.

In Grade 1, the AUCs for the mCLASS Lectura Composite Score ranged from .82 to .89, suggesting again that it is a strong predictor of the SELSp scale score. Sensitivity for the composite score ranged from .67

to .70, while its specificity ranged from .78 to .81. FNL, LSS, FEP, and FLO_WRC all had most of their AUCs above .80. FSS, similar to its performance in kindergarten, had relatively low AUCs. This might be due to the fact that SELSp in Grade 1 did not contain a component measuring phonological awareness and/or that the FSS subtest had low validity and reliability. Overall, sensitivity and specificity ranged from .70 to .80.

In Grade 2, the AUCs for the mCLASS Lectura Composite Score ranged from .84 to .87, sensitivity ranged from .61 to .79, and specificity ranged from .76 to .83. AUCs for FEP and FLO_WRC were above .80; FLO_ACC had lower AUCs, but still close to .80. CP had the lowest AUCs on average compared with other subtests, indicating that it was not very accurate in predicting students' performance on SELSp.

In Grade 3, the AUCs for the mCLASS Lectura Composite Score ranged from .82 to .92, its sensitivity ranged from .74 to .78, and its specificity ranged from .76 to .85. FEP and FLO_WRC continued to be the strongest predictors of students' performance on SELSp, as can be seen from the high AUCs. FLO_ACC and CP had lower AUCs on average but still remained robust predictors. The average sensitivity of the subtests was .76; and the average specificity was .78.

In Grades 4–5, FLO_WRC and CP were the strongest predictors of students' performance on SRSp, which might be due to SRSp having a strong component measuring reading comprehension. FLO_ACC had the weakest predictive power, which again could be explained by the fact that as more students in these upper grades were able to obtain a high score on reading accuracy, the power of this test to differentiate among students with different levels of reading skills also diminished.

The AUCs for the mCLASS Lectura Composite Score and all subtests for all grade levels and TOYs are presented in Tables 1–6 in the Appendix. Overall, the AUCs for the mCLASS Lectura Composite Score across TOY for kindergarten through Grade 3 are greater than .80. Of the 24 predictive models between mCLASS Lectura Composite Scores and WM-AP raw scores in kindergarten and SELSp scaled scores in Grades 1–3, interpretation of the AUC values suggests moderate predictive models for approximately one-third (29.17%) of the models, with AUCs ranging from .75–.84, and very good predictive models for 70.83% of the models (AUCs ranging from .85–.95; Smolkowski & Cummings, 2015; Swets, 1988). In Grade 4, of the six predictive models between mCLASS Lectura Composite Scores and SRSp scaled scores, interpretation of the AUC values suggests moderate predictive models for each TOY with AUCs ranging from .78 to .82. Finally, in Grade 5, of the six predictive models between mCLASS Lectura Composite Scores and SRSp scaled scores, interpretation of the AUC values suggest poor predictive models for half of the models, with AUCs ranging from .61–.74, and moderate predictive models for half of the models, with AUCs ranging from .75–.79.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6), 716–723.

Alonzo, J., Gonzalez, M., & Tindal, G. (2013). *The Development of easyCBM Spanish Literacy Assessments for Use in Grades K-2*. (Technical Report No. 1301). University of Oregon, Behavioral Research and Teaching.

American Educational Research Association, National Council on Measurement in Education, & American Psychological Association [AERA, NCME, & APA] (2014). *The Standards for educational and psychological testing*. American Educational Research Association.

Anthony, J. L., Williams, J. M., McDonald, R., Corbitt-Shindler, D., Carlson, C. D., & Francis, D. J. (2006). Phonological processing and emergent literacy in Spanish-speaking preschool children. *Annals of Dyslexia*, 56(2), 239–270.

August, D., & Shanahan, T. (2006). *Developing literacy in second-language learners: Report of the national literacy panel on language minority children and youth*. Lawrence Erlbaum Associates.

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin*, 107(2), 238–246.

Bollen, K. A. (1989). *Structural equations with latent variables*. Wiley.

Browne, M. W., & Cudeck, R. (1992). Alternative ways of assessing model fit. *Sociological Methods & Research*, 21(2), 230–258.

Brownell, R. (2012). *Expressive One Word Picture Vocabulary Test - bilingual version* [Measurement instrument]. Riverside.

Christ, T. J., & Ardoin, S. P. (2009). Curriculum-based measurement of oral reading: Passage equivalence and probe-set development. *Journal of School Psychology*, 47, 55–75. https://doi.org/10.1016/jjsp.2008.09.004.

Cicchetti D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment*, *6*(4), 284–290.

Crawford, A. (1985). Fórmula y gráfico para determinar la comprensibilidad de textos del nivel primario en castellano. *Lectura Y Vida*, 4, 18–24.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16(3), 297–334.

Davies, M., & Davies, K. H. (2018). *A frequency dictionary of Spanish: Core vocabulary for learners* (2nd ed.). Routledge.

Deno, S. (1992). The nature and development of curriculum-based measurement. *Preventing School Failure*, 36(2), 5–10.

López-Escribano, C., Elosúa de Juan, M. R., Gómez-Veiga, I., & García-Madruga, J. A. (2013). A predictive study of reading comprehension in third-grade Spanish students. *Psicothema*, 25(2), 199–205. https://doi.org/10.7334/psicothema2012.175

Habibzadeh, F., Habibzadeh, P., & Yadollahie, M. (2016). On determining the most appropriate test cut-off value: The case of tests with continuous results. *Biochemia Medica*, 26(3), 297–307.

Hammer, C. S., Miccio, A. W., & Rodriguez, B. L. (2012). Bilingual language acquisition and the child socialization process. In B. A. Goldstein (Ed.) *Bilingual language development and disorders in Spanish-English speakers* (2nd ed., pp. 31–46). Paul H. Brookes.

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: a Multidisciplinary Journal*, *6*(1), 1–55.

Imagination Station. (2016). Istation's Indicators of Progress Español Technical Report. Author.

Jenkins, J.R., Fuchs, L. S., van den Broek, P., Espin, C., & Deno, S. L. (2003). Sources of individual differences in reading comprehension and reading fluency. *Journal of Educational Psychology*, *95*(4), 719–729.

Jenkins, J. R., Hudson, R. F., & Johnson, E. S. (2007). Screening for at-risk readers in a response to intervention framework. *School Psychology Review*, 36(4), 582–600. https://doi.org/10.1080/02796015.2007.12087919

Katzir, T., Hershko, S., & Halamish, V. (2013). The effect of font size on reading comprehension on second and fifth grade children: Bigger is not always better. *PLoS ONE*, *8*(9), e74061. https://doi.org/10.1371/journal.pone.0074061.

Kenny, D. A., Kaniskan, B., & McCoach, D. B. (2015). The performance of RMSEA in models with small degrees of freedom. *Sociological Methods & Research*, 44(3), 486–507.

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, *15*(2), 155–163.

Jackson-Maldonado et al. (2003). *MacArthur-Bates communication development inventory in Spanish*. Paul H. Brookes.

Miguez-Álvarez, C., Cuevas-Alonso, M., & Saavedra, Á. (2021). Relationships Between Phonological Awareness and Reading in Spanish: A Meta-Analysis. *Language Learning*. https://doi.org/10.1111/lang.12471

National Center on Intensive Intervention. (2018). *Academic screening tools chart rating rubric* Retrieved October 13, 2022 from https://intensiveintervention.org/sites/default/files/NCII_AcademicScreening_RatingRubric_July2018.pdf

Peña, E. D., Kester, E. S., & Sheng, L. (2012). Semantic development in Spanish-English bilinguals: Theory, assessment, and intervention. In B. Goldstein (Eds.), *Bilingual language development & disorders* (pp. 131–152). Paul H. Brookes.

Pepe, M. S. (2003). *The statistical evaluation of medical tests for classification and prediction*. Oxford University.

Powell-Smith, K. A. Good, R. H. & Atkins, T. (2010). DIBELS Next Oral Reading Fluency Readability Study (Technical Report No. 7). Dynamic Measurement Group.

Renaissance Learning, Inc. (2021). S*tar Assessments for Spanish - Early Literacy Technical Documentation*. Author.

Renaissance Learning, Inc. (2018). *Star Assessments for Spanish - Reading Technical Documentation*. Author.

Renaissance Learning, Inc. (2014). *Star Assessments for Spanish - Reading Technical Documentation*. Author.

Salvia, J., Ysseldyke, J., & Witmer, S. (2016). *Assessment in Special and Inclusive Education* (13th ed.). Cengage.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 461–464.

Smolkowski, K., & Cummings, K. D. (2015). Evaluation of diagnostic systems: The selection of students at risk of academic difficulties. *Assessment for Effective Intervention*, 41(1), 41–54. https://doi.org/10.1177/1534508415590386

Spanish language arts curriculum: Maravillas Wonders. McGraw Hill. (n.d.). Retrieved April 14, 2022, from https://www.mheducation.com/prek-12/program/microsites/MKTSP-BGA10M0/maravillas.html

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

U. S. Department of Commerce (2020). Census regions and divisions of the United States. *United States Census Bureau, Geography Division*.

Woodcock, R., Alvarado, C. G., Schank, F. A., Mather, N., Wendling, B., & Muñoz-Sandoval, A. F. (2017). *Woodcock-Muñoz Test of Achievement [Batería Woodcock-Muñoz pruebas de aprovechamiento]* (4th ed.). [Measurement instrument]. Riverside Insights.

Appendix

mCLASS Lectura Cut Scores

Table A1: Cut Scores and Receiver Operator Characteristic (ROC) Curve Results for mCLASS Lectura Kindergarten

Measure	Criterion	TOY	Cut Score	Sensitivity	Specificity	AUC	95% CI
CS	20th	BOY	255	.73	.83	.85	.79, .92
		MOY	334	.78	.92	.95	.92, .98
		EOY	370	.80	.93	.96	.93, .98
	40th	BOY	268	.77	.71	.80	.73, .86
		MOY	351	.74	.79	.85	.80, .91
		EOY	388	.76	.83	.86	.81, .92
FNL	20th	BOY	2	.69	.74	.78	.71, .85
		MOY	13	.91	.80	.91	.87, .96
		EOY	24	.84	.83	.94	.91, .98
	40th	BOY	5	.83	.62	.77	.70, .84
		MOY	19	.76	.70	.82	.76, .89
		EOY	29	.68	.72	.82	.76, .89
FSS	20th	BOY	14	.76	.70	.76	.68, .83
		MOY	29	.67	.82	.84	.74, .89
		EOY	38	.44	.83	.72	.63, .81
	40th	BOY	22	.74	.50	.67	.59, .74
		MOY	33	.59	.75	.69	.61, .77
		EOY	43	.47	.77	.61	.53, .70
FSL	20th	BOY	3	.67	.76	.79	.72, .87
		MOY	15	.82	.82	.90	.84, .95
		EOY	24	.80	.85	.92	.88, .96
	40th	BOY	5	.72	.72	.79	.72, .86
		MOY	21	.76	.71	.81	.74, .87
		EOY	29	.72	.77	.84	.78, .90
LSS	20th	BOY	1	.96	.29	.31	.26, .36
		MOY	5	.87	.71	.84	.78, .90
		EOY	11	.87	.82	.93	.90, .97
	40th	BOY	1	.92	.34	.70	.64, .76
		MOY	7	.82	.70	.84	.78, .89
		EOY	18	.86	.71	.87	.82, .92

Measure	Criterion	TOY	Cut Score	Sensitivity	Specificity	AUC	95% CI
FEP	20th	BOY	1	1.0	.28	.76	.70, .82
		MOY	3	.91	.65	.90	.86, .95
		EOY	5	.84	.85	.93	.88, .97
	40th	BOY	1	.91	.30	.71	.64, .78
		MOY	4	.86	.66	.86	.81, .92
		EOY	10	.83	.79	.89	.85, .94

Table A2: Cut Scores and Receiver Operator Characteristic (ROC) Curve Results for mCLASS Lectura Grade 1

Measure	Criterion	TOY	Cut Score	Sensitivity	Specificity	AUC	95% CI
CS	20th	BOY	337	.70	.81	.86	.82, .89
		MOY	376	.70	.88	.89	.86, .92
		EOY	419	.70	.87	.88	.85, .91
	40th	BOY	347	.67	.78	.82	.79, .86
		MOY	389	.70	.83	.85	.82, .88
		EOY	433	.67	.86	.84	.81, .87
FNL	20th	BOY	23	.67	.81	.80	.76, .85
		MOY	30	.67	.83	.85	.81, .88
		EOY	36	.69	.78	.80	.76, .84
	40th	BOY	26	.58	.82	.76	.72, .80
		MOY	34	.61	.82	.78	.74, .82
		EOY	39	.63	.73	.75	.70, .79
FSS	20th	BOY	25	.53	.69	.69	.64, .73
		MOY	34	.36	.85	.67	.62, .72
		EOY	40	.41	.76	.64	.59, .69
	40th	BOY	31	.61	.59	.62	.57, .67
		MOY	40	.48	.66	.61	.56, .66
		EOY	47	.51	.63	.60	.56, .65
FSL	20th	BOY	21	.59	.79	.79	.75, .83
		MOY	30	.54	.83	.79	.74, .83
		EOY	37	.52	.83	.75	.70, .80
	40th	BOY	25	.58	.74	.74	.70, .78
		MOY	35	.56	.75	.71	.67, .76
		EOY	41	.50	.74	.68	.63, .72

Measure	Criterion	TOY	Cut Score	Sensitivity	Specificity	AUC	95% CI
LSS	20th	BOY	10	.76	.72	.82	.79, .86
		MOY	20	.80	.80	.87	.84, .91
		EOY	29	.72	.82	.86	.82, .89
	40th	BOY	13	.71	.72	.80	.77, .84
		MOY	24	.74	.80	.84	.81, .87
		EOY	35	.69	.78	.82	.79, .86
FEP	20th	BOY	5	.83	.71	.83	.79, .86
		MOY	11	.72	.84	.87	.84, .91
		EOY	20	.69	.85	.87	.84, .90
	40th	BOY	7	.75	.73	.81	.77, .85
		MOY	17	.73	.78	.84	.81, .88
		EOY	27	.73	.79	.84	.81, .88
FLO_WRC	20th	BOY	5	.77	.71	.84	.80, .88
		MOY	10	.72	.79	.87	.83, .90
		EOY	19	.72	.84	.87	.84, .91
	40th	BOY	7	.71	.71	.81	.78, .85
		MOY	14	.72	.78	.84	.81, .87
		EOY	27	.68	.85	.85	.81, .88
FLO_ACC	20th	BOY	40	.74	.75	.82	.78, .86
		MOY	66	.70	.80	.84	.80, .88
		EOY	81	.71	.81	.84	.80, .88
	40th	BOY	60	.76	.67	.78	.74, .82
		MOY	81	.76	.67	.80	.77, .84
		EOY	91	.73	.67	.79	.75, .83

Table A3: Cut Scores and Receiver Operator Characteristic (ROC) Curve Results for mCLASS Lectura Grade 2

Measure	Criterion	TOY	Cut Score	Sensitivity	Specificity	AUC	95% CI
CS	20th	BOY	329	.70	.75	.83	.77, .90
		MOY	369	.73	.80	.87	.81, .93
		EOY	408	.61	.83	.84	.77, .91
	40th	BOY	340	.72	.76	.84	.79, .88
		MOY	387	.78	.76	.85	.81, .90
		EOY	429	.79	.77	.86	.82, .90
FEP	20th	BOY	9	.70	.75	.82	.76, .89
		MOY	15	.79	.78	.87	.82, .93
		EOY	20	.64	.82	.85	.78, .92
	40th	BOY	14	.73	.75	.82	.78, .87
		MOY	21	.74	.77	.84	.80, .89
		EOY	27	.71	.81	.86	.82, .90
FLO_WRC	20th	BOY	18	.67	.74	.83	.77, .90
		MOY	32	.76	.79	.86	.80, .92
		EOY	34	.61	.84	.83	.76, .90
	40th	BOY	25	.72	.78	.84	.80, .88
		MOY	47	.77	.77	.85	.81, .89
		EOY	53	.79	.76	.86	.82, .90
FLO_ACC	20th	BOY	80	.73	.71	.81	.72, .89
		MOY	90	.76	.76	.82	.72, .91
		EOY	90	.70	.82	.78	.66, .89
	40th	BOY	90	.76	.70	.78	.72, .84
		MOY	95	.72	.67	.79	.73, .84
		EOY	95	.65	.76	.73	.67, .80
CP	20th	BOY	0.5	.58	.60	.67	.59, .75
		MOY	1	.64	.67	.73	.66, .80
		EOY	1.5	.70	.61	.70	.62, .78
	40th	BOY	1.5	.75	.56	.66	.61, .72
		MOY	2.5	.78	.61	.74	.69, .79
		EOY	3	.83	.62	.79	.75, .84

Table A4: Cut Scores and Receiver Operator Characteristic (ROC) Curve Results for mCLASS Lectura Grade 3

Measure	Criterion	TOY	Cut Score	Sensitivity	Specificity	AUC	95% CI
CS	20th	BOY	329	.78	.82	.92	.85, .99
		MOY	370	.78	.85	.89	.79, .99
		EOY	407	.78	.86	.93	.85, 1.0
	40th	BOY	343	.77	.76	.82	.76, .89
		MOY	390	.74	.77	.85	.79, .90
		EOY	431	.79	.79	.89	.85, .94
FEP	20th	BOY	13	.78	.81	.90	.80, .99
		MOY	17	.78	.86	.88	.76, 1.0
		EOY	20	.78	.86	.92	.82, 1.0
	40th	BOY	19	.79	.71	.80	.73, .87
		MOY	26	.74	.75	.83	.77, .89
		EOY	31	.79	.73	.85	.79, .91
FLO_WRC	20th	BOY	29	.89	.80	.92	.86, .98
		MOY	42	.78	.85	.89	.79, .98
		EOY	60	.78	.86	.92	.84, 1.0
	40th	BOY	36	.77	.74	.82	.76, .89
		MOY	55	.74	.77	.84	.79, .90
		EOY	78	.81	.79	.89	.84, .94
FLO_ACC	20th	BOY	90	.89	.66	.91	.83, .99
		MOY	90	.67	.88	.84	.66, 1.0
		EOY	90	.78	.93	.87	.68, 1.0
	40th	BOY	95	.81	.41	.74	.65, .83
		MOY	95	.70	.76	.76	.68, .84
		EOY	95	.65	.81	.76	.67, .85
CP	20th	BOY	1.5	.78	.86	.87	.74, .94
		MOY	2	.89	.71	.80	.68, .93
		EOY	2.5	.56	.76	.80	.68, .92
	40th	BOY	2	.56	.86	.80	.73, .87
		MOY	2.5	.74	.72	.78	.71, .85
		EOY	3.5	.72	.77	.81	.75, .87

Table A5: Cut Scores and Receiver Operator Characteristic (ROC) Curve Results for mCLASS Lectura Grade 4

Measure	Criterion	TOY	Cut Score	Sensitivity	Specificity	AUC	95% CI
CS	20th	BOY	340	.62	.80	.80	.74, .85
		MOY	389	.62	.83	.80	.75, .86
		EOY	422	.60	.83	.81	.75, .86
	40th	BOY	356	.64	.81	.80	.74, .85
		MOY	397	.62	.89	.82	.77, .87
		EOY	434	.61	.83	.78	.72, .83
FLO_WRC	20th	BOY	51	.66	.78	.80	.74, .85
		MOY	60	.62	.82	.80	.74, .86
		EOY	63	.60	.81	.80	.74, .86
	40th	BOY	60	.66	.81	.80	.74, .85
		MOY	65	.61	.89	.82	.76, .87
		EOY	70	.61	.82	.77	.71, .83
FLO_ACC	20th	BOY	90	.51	.83	.73	.66, .80
		MOY	90	.26	.95	.67	.59, .74
		EOY	90	.29	.94	.67	.60, .74
	40th	BOY	95	.70	.64	.74	.67, .80
		MOY	95	.37	.91	.67	.61, .74
		EOY	95	.49	.77	.69	.63, .76
CP	20th	BOY	3	.47	.77	.71	.64, .78
		MOY	4	.62	.84	.80	.74, .86
		EOY	4.5	.58	.88	.83	.78, .88
	40th	BOY	4.5	.57	.64	.68	.61, .74
		MOY	5.5	.57	.80	.78	.72, .84
		EOY	7	.63	.90	.82	.77, .87

Table A6: Cut Scores and Receiver Operator Characteristic (ROC) Curve Results for mCLASS Lectura Grade 5

Measure	Criterion	TOY	Cut Score	Sensitivity	Specificity	AUC	95% CI
CS	20th	BOY	360	.58	.71	.73	.67, .80
		MOY	383	.60	.80	.77	.71, .84
		EOY	427	.61	.74	.79	.73, .85
	40th	BOY	364	.57	.70	.71	.65, .78
		MOY	391	.61	.74	.73	.66, .79
		EOY	435	.61	.72	.75	.69, .81
FLO_WRC	20th	BOY	70	.58	.69	.73	.66, .79
		MOY	83	.60	.79	.77	.70, .83
		EOY	84	.61	.74	.78	.72, .84
	40th	BOY	72	.57	.70	.71	.65, .77
		MOY	89	.60	.74	.72	.66, .78
		EOY	91	.61	.71	.74	.68, .80
FLO_ACC	20th	BOY	90	.23	.84	.60	.52, .68
		MOY	90	.14	.98	.66	.58, .74
		EOY	90	.18	.97	.68	.61, .75
	40th	BOY	95	.53	.68	.66	.59, .72
		MOY	95	.32	.92	.64	.57, .70
		EOY	95	.32	.85	.66	.60, .73
CP	20th	BOY	4	.57	.79	.74	.67, .81
		MOY	4.5	.55	.80	.78	.72, .83
		EOY	5	.65	.79	.78	.72, .84
	40th	BOY	4.5	.50	.75	.69	.63, .75
		MOY	5.5	.61	.78	.76	.71, .82
		EOY	6.5	.61	.80	.80	.74, .85