Predictive Validity of Literably Oral Reading Fluency

Technical Report No. 1

Joseph B. Townsend[1]

Benjamin W. Domingue[2]

December 2018

Introduction

This report provides information on the technical properties of the Literably K-8 reading assessment. Using student data from two districts, one urban and one rural, we present findings related to the validity of the Literably literacy assessment, focusing, in particular, on the criterion validity and the classification accuracy of the assessment. This is the first report detailing the validity properties of the Literably screening tool. The findings suggest that the technical characteristics of Literably scores are both (1) in line with other early literacy screening tools and (2) adequate for the screening of young readers.

The Literably literacy assessment for grades K-8 is designed to inform instruction, screen students for reading difficulties, identify students' reading levels, and monitor students' response to intervention. The Literably assessment is similar to other early literacy screening tools in that students are scored on the number of words they read out loud correctly in one minute; we denote this as the words correct per minute score (WCPM), although another common term is oral reading fluency (ORF). Literably, however, differs from many of the older reading assessments in how it is administered. In lieu of reading out loud to a teacher or instructor, students in the Literably system read to a tablet or computer. While this change has the potential to save educators valuable classroom time, such time gains would be negated if the validity of the assessment is not on par with traditional screening tools.

[1]Joseph B. Townsend, PhD, is a graduate of the Stanford Graduate School of Education. His contribution to this publication was as a paid consultant to Literably, Inc., and was not part of his Stanford University duties or responsibilities.

[2]Benjamin W. Domingue, PhD, is an Assistant Professor at the Stanford Graduate School of Education. His contribution to this publication was as a paid consultant to Literably, Inc., and was not part of his Stanford University duties or responsibilities.

We focus on two types of measures in this report: criterion validity and classification accuracy. Criterion validity is the extent to which performance on a criterion measure can be estimated based on performance on the assessment being validated (Salvia, Ysseldyke, and Witmer 2012). To estimate criterion validity, we present results of predictive validity analyses, correlating Literably scores with end-of-year Smarter Balanced Assessment Consortium (SBAC) English Language Arts (ELA) scores. Classification accuracy refers to how well an assessment distinguishes between two groups of students at different skill levels. For our purposes, we examine classification accuracy by evaluating how well Literably distinguishes between student who will meet the SBAC ELA achievement standard as compared to those who will not.

While results vary both between- and within-district across grades, we find that the predictive validity of the Literably assessment is commensurate with other, well-researched and widely used early literacy screening tools such as DIBELS™, AIMSweb™, and easyCBM™. Classification accuracy results are similarly encouraging, as these values are within the same ranges reported by these assessments as well. These early results suggest that Literably is adequate for universal screening in English Language Arts in the elementary grades.

Literably Literacy Assessment

Literably is a reading assessment for K-8 students designed to help educators identify reading levels, diagnose skill gaps, screen students for reading difficulties, and monitor reading learning progress. The assessment consists of students reading passages out loud and then answering a series of comprehension questions. While the administration of the assessment may differ from other early literacy screening tools, such as DIBELS™, AIMSweb™, and easyCBM™, Literably is generally analogous to these assessments.

A student using Literably receives an oral reading fluency (ORF), or a words correct per minute (WCPM) score based on the reading out loud portion of the assessment. These scores measure the number of words a student reads out loud correctly in one minute. Research on the diagnostic accuracy of WCPM has supported the use of these scores for the purpose of screening elementary school students for reading difficulties (Kilgus et al. 2014). In the analyses that follow, we utilize these WCPM scores to measure the predictive criterion validity and classification accuracy of Literably scores.

Passages

Literably's passages are intended to resemble the materials that students read in the classroom. As such, Literably draws its reading passages from leveled children's trade books. Approximately half are excerpts from books, while the remainder are full books. Roughly half are fiction and half non-fiction. Pictures are available for all passages up through the 3rd grade level. Literably reports that all books were officially leveled according to Heinemann's Guided Reading framework (Fountas and Pinnell 1996).

Administration

Literably is administered using a computer or tablet. Instead of reading out loud to an educator, students taking a Literably assessment read out loud to an application that digitally records the students reading. This differs markedly from traditional early reading screening instruments, which nearly all involve a teacher or instructor administering the assessment to a student one-on-one, and scoring the test in real time.

Scoring

Recordings of students reading are provided to Literably-hired raters, who transcribe the audio recordings. WCPM scores are computed from these transcriptions, and then made available to schools, educators, and parents through the Literably application. Raters face random test recordings to ensure that their transcriptions show a high rate of concordance with the transcriptions of Literably staff members. For more on raters, see Literably Reliability Report, 2018.

Method

Criterion Validity

To estimate criterion validity, we correlate Literably WCPM scores with end-of-year Smarter Balanced Assessment Consortium (SBAC) English Language Arts (ELA) raw scores. This analysis is conducted for grades 2-5 in each season (fall, winter, and spring corresponding to the beginning, middle, and end of the school year) for both school districts.

A robust body of literature examines criterion validity for early reading assessments using the correlation approach employed here (e.g., L. S. Fuchs et al. 2001; Good, Simmons, and Kame'enui 2001; Buck and Torgesen 2003; Danne et al. 2005; Riedel 2007; Good et al. 2018). As we do in this report, many of these studies correlate ORF scores with state-mandated ELA and Reading assessments, which are often taken weeks or months in the future, while others studies correlate ORF scores with concurrently administered ELA or Reading assessments. The correlations reported in these studies vary both by grade level, by assessment type, and by assessment time of year. In general, publisher-reported results from correlations of ORF scores with end-of-year reading assessments fall in the range of .60 to .75 (NCS Pearson 2012; Good et al. 2013; Christ and Colleagues 2015; Good et al. 2018).

Students often take the Literably assessment multiple times within a season. For the purposes of measuring criterion validity, only one score per student is needed per season. We select the first non-zero Literably WCPM score within the district's official screening window, which typically falls in the beginning of the season.

We focus on correlations for grades 3-5 but also present correlations for grade 2. Students in second grade, however, do not take the end-of-year SBAC assessment. Findings, then, are the result of correlating second grade WCPM scores with grade 3 SBAC scores. Given the increased length of time elapsed between the two measures, we expect some decrease in the grade 2 correlations.

Classification Accuracy

We take a variety of approaches to estimating classification accuracy. In general, the goal of these analyses is to use WCPM scores to predict success on an external criterion. The criterion used here is meeting the state-defined proficiency standard on the SBAC ELA and Reading assessments; a student meets a proficiency standard if they score above a cutpoint predetermined by the state. A minimal standard for a screening tool is that, more often than not, it successfully distinguishes between students who will meet the standard versus those who will not. A more useful screening tool will have higher levels of classification accuracy; below we discuss various metrics for determining such accuracy.

The first approach reported below is the area under the receiver operating characteristic curve (ROC curve). ROC curve analyses have become common when evaluating the classification accuracy of assessments, and are often used to assist in setting benchmark, or cut scores, when using an assessment to predict a dichotomous outcome on an external criterion (e.g., Riedel 2007; NCS Pearson 2012; Christ and Colleagues 2015). High ROC scores indicate that

the screening tool has strong classification accuracy. Such scores can be interpreted in the following way: a given ROC score indicates the probability that a randomly selected student meeting the standard will have a higher score than a randomly selected student not meeting the standard. In general, scores above .90 are considered excellent; scores between .80 and .90 are consider good, and scores less than .69 are considered poor (Christ and Colleagues 2015). A ROC score of 0.5 indicates that the tool does not successfully distinguish between students who do and do not meet the standard (e.g., whose score is higher is effectively a coin toss). A survey of publisher-reported ROC curve analyses indicates that ORF-based screening tools often have area under the ROC curve values in the .70 to .95 range (Riedel 2007; NCS Pearson 2012; Christ and Colleagues 2015).

Results from logistic regressions are also reported; we focus on Nagelkerke's R^2. While R^2 values have a variety of interpretations, Nagelkerke's R^2 may generally be considered a measure of how well screening scores explain variation in the criterion outcome. For our purposes, then, WCPM scores are used to predict successfully meeting the SBAC ELA and Reading standard. Once again, higher scores–those closer to 1.0—indicate a stronger screening tool. R^2 values in the range of .40 to .55 are common for elementary reading assessments (Good et al. 2018). Note that these values do not have the same intuitive interpretation as the ROC scores.

In addition to these two general measures of classification performance, we also present descriptive statistics on how Literably "cut scores" screened students. Literably provides school district clients with benchmark scores that signal, for students achieving at or above the benchmark, a high likelihood of meeting the end-of-year standard. We examine three statistics related to a district's use of these cut scores: *sensitivity*, *specificity*, and *percent correct*. Note that the ROC analyses discussed above also incorporate information about sensitivity and specificity, but do so in an effort to evaluate the classification accuracy of the assessment as a whole. The sensitivity and specificity values we report are related to a single, specific cut score.

Sensitivity measures the true positive rate, which is defined as the true positive cases divided by the sum of true positive and false negative cases. A true positive case is one in which a student has a WCPM score below the benchmark and indeed fails to meet standard on the criterion. A false negative case occurs when a student scores above the benchmark score but does not achieve the goal.

Specificity, or the true negative rate, is defined as the true negative cases divided by the sum of the true negative cases and the false positive cases. A true negative is one in which a

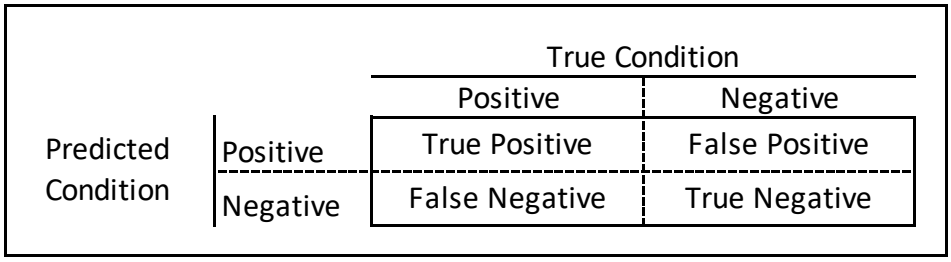

Figure 1: Contingency Table Showing True Positive, False Postive, False Negative, and True Negative Outcomes

student scored above the benchmark and also meets standard on the end-of-year assessment. A false positive is when a student scores below the benchmark but does meet standard.

Figure 1 shows the outcomes true positive, false positive, false negative, and true negative as the cross between the predicted outcome, and the true condition.

Finally, the percent correct is simply the proportion of cases correctly classified using the Literably-provided cut score. See Setting Literably Cut Scores (Townsend and Domingue 2018) for more on how Literably set the cut scores.

Data

Data for this validity study come from two California school districts. The first district (District A) is a mid-sized urban school district enrolling a diverse student population in grades pre-K through 8. The second district (District B) is a small, rural district, which also enrolls students in grades pre-K through 8. District A provided data for two academic school years, 2015-2016, and 2016-2017, while District B provided data for only the 2016-2017 academic year. Both districts provided student-level data, including grade, English Language Learner (ELL) status, end-of-year SBAC ELA scores, and SBAC ELA and Reading performance levels. Each student record included an anonymized student ID linkable to Literably records.

Literably also provided student-level records of performance on the Literably assessment, in the form of WCPM scores. Students often take the Literably assessment multiple times per year, with many students taking it multiple times per season. For all analyses presented in this report, we select only one score per season. Literably provided a date range for each district and time of year where students were likely to be taking the test for screening purposes. We selected the first non-zero WCPM score from this range for each student. We select the first WCPM score in this range, as opposed to a later score or some combination

of scores, as this is likely to result in the most conservative - that is, the lowest - observed correlation between WCPM scores and end-of-year SBAC scores. WCPM scores later in the range, and therefore closer to the end-of-year assessment, may capture actual changes in a student's reading ability, which will also likely be reflected in end-of-year SBAC scores; this would drive up the correlation between WCPM and SBAC scores, relative to earlier WCPM scores.

Harder Literably reading passages are expected to yield lower WCPM scores. Since Literably passages vary in difficulty (even amongst passages at the same grade level), this means that a student's WCPM score is expected to vary across alternate forms of the assessment. To address this, Literably also provided equated scores, or WCPM scores designed to be comparable across season, within grade. The validity results presented below use the equated version of the selected score, when available (for more on score equating, see Literably Technical Manual 2018). All WCPM records were linked to the school district records using the anonymized student IDs.

Results

Sample Description

Table 1 reports basic descriptive information on Districts A and B for the 2016-2017 school year. The tables include counts by grade, percent ELL, mean and standard deviation for SBAC ELA scores, and counts of students meeting or surpassing the SBAC ELA standard.

We opt to present these, and the majority of results that follow, for a pooled sample of District A and District B; in many cases we also include the data from District A for the 2015-2016 school year. We pool the data to protect the anonymity of the two districts participating in the study. For this same reason, we do not present descriptive or summary statistics for District A alone for the 2015-2016 school year.

While we seek to protect the anonymity of each district, we can report that District A is more than twice the size of District B at each grade level. District A also enrolls a higher percentage of ELL students than does District B.

The pooled, 2016-2017 sample shows that, generally, about 40% of the sample is ELL. This is about double the rate of the average California public school.[3] The mean ELA scores are

[3]See CalEdFacts for more.

higher than the state average, at each grade.[4]

Table 1: Descriptive Statistics for Districts A and B, 2016-2017

Grade	N	Pct. ELL	Mean ELA Score	SD ELA Score	Pct. Meeting Standard
3	635	40	2462.5	86.5	65
4	698	36	2508.5	93.7	68
5	637	41	2541.5	103.0	65

Table 2 presents WCPM statistics for the full pooled sample, including all data from District A and District B. Given the centrality of language to these analyses, it also includes a breakdown by ELL status. For nearly each grade and group combination, we see that WCPM scores increase within grade across times of year. We also observe that WCPM scores increase grade-to-grade, a pattern most notable when comparing Fall scores across grades.

There are, however, potentially important exceptions to these patterns. Notice, for example, that non-ELL students in grades 4 and 5, show mean WCPM scores around 110 in both the middle and end of year. This is notable because we would expect to observe a seasonal increase in mean scores from winter to spring, and a grade-over grade increase as well. That WCPM scores level out at about 110 suggests there may be a ceiling effect - that is, students cannot consistently score above this, regardless of their reading level. This risks the possibility that, for students with relatively high reading ability, the WCPM score is not differentiating students effectively. If that is indeed the case, we would expect to observe lower correlations with end-of-year SBAC scores.

Table 2 also makes clear that ELL and non-ELL students in the same grade do perform differently on the Literably assessment. Non-ELL students have higher WCPM than same-grade and season ELL students in all cases. The difference is roughly 10 words in grades 2 and 3, and grows to a difference of roughly 12-15 in grades 4 and 5.

[4]See California Assessment of Student Performance and Progress for more.

Table 2: Summary Statistics for Literably Words Correct per Minute for Districts A and B, 2015-2017

	Grade 2			Grade 3			Grade 4			Grade 5		
Season	N	Mean	SD	N	Mean	SD	N	Mean	SD	N	Mean	SD
All												
Fall	373	65.1	30.3	1107	76.8	35.7	1183	98.1	32.2	1081	103.0	33.1
Winter	383	73.6	28.9	1110	81.7	34.3	1202	105.4	32.2	1062	107.3	31.7
Spring	381	80.6	26.0	1037	85.5	33.7	1192	107.6	29.7	1035	108.0	29.8
ELL												
Fall	181	61.0	30.9	238	61.3	33.4	230	88.0	33.8	257	93.2	33.7
Winter	186	67.8	28.1	243	70.2	35.9	240	95.8	31.1	250	94.9	30.5
Spring	186	75.8	26.6	229	78.0	32.9	240	97.9	27.5	235	94.8	29.6
Non-ELL												
Fall	192	69.1	29.2	362	70.8	34.7	415	103.0	30.4	364	106.8	33.8
Winter	197	79.0	28.7	371	83.5	33.6	425	110.4	30.0	335	110.1	29.9
Spring	195	85.2	24.7	349	89.7	30.3	429	109.1	27.8	325	109.0	27.1

Criterion Validity

We estimate predictive criterion validity by correlating student WCPM with raw, end-of-year SBAC ELA scores.

Table 3 reports these correlations for Districts A and B during the 2015-2016, and 2016-2017 academic years. This table includes results for students in grade 2, although these students do not take the SBAC assessment. This result comes from correlating grade 2 WCPM scores with these students' third grade SBAC performance.

Results for District A students in grades 3-5, for both the 2015-2016 and 2016-2017 academic years, show correlations that are generally in the .60-.70 range. The results tend to be somewhat weaker in the higher grades, where we observe a handful of correlations in the .50-.60 range. Recall that this is also where we observed a potential ceiling to the WCPM scores. Clearly these correlations vary within- and across-grade.

Results for District B, grades 3-5, are analogous to those from District A: the correlations generally fall in the .60-.70 range, although there are two sub .60 results found in grade 5.

We expected to see weaker correlations for second grade relative to grades 3-5, due to the lag in taking the external criterion. In fact, we do not observe this pattern. One explanation may be survivor bias: students who successfully get promoted from grade 2 to grade 3 and those who remain in the district are likely to differ in crucial ways from those who do not. The students who persist in District A from grade 2 through the end of grade 3 are likely to be higher achieving students, on average, than the full population of students who begin grade 2, which includes students who fail to get promoted at the end of the year, and students who may leave the district. Higher achieving students may show a stronger correlation between their Literably WCPM and SBAC ELA scores.

Finally, Table 3 also reports results from both districts combined. Here we see more stability in the correlations, and also slightly higher overall values. Only in the spring of grade 5 does the correlation fall below .60, and only once.

As noted previously, the leveling off of WCPM scores in the spring of 4th and 5th grade may represent a ceiling effect: students in these grades, regardless of their reading ability, cannot consistently score above this threshold. This raises the possibility that WCPM scores may not be differentiating between students with relatively high reading abilities. The low correlations observed in 5th grade between WCPM scores and SBAC ELA scores lend support this hypothesis.

In sum, these results are commensurate with other, well-researched early literacy screening tools (NCS Pearson 2012; Good et al. 2013; Christ and Colleagues 2015).

Table 3: Criterion Validity for Literably WCPM Score, Districts A and B, 2015-2017

	Grade 2		Grade 3		Grade 4		Grade 5	
Season	N	Corr.	N	Corr.	N	Corr.	N	Corr.
District A: 2015-16								
Fall	373	0.65	507	0.63	538	0.60	460	0.52
Winter	383	0.65	496	0.67	537	0.70	477	0.61
Spring	381	0.69	459	0.74	523	0.73	475	0.60
District A: 2016-17								
Fall			460	0.66	483	0.70	464	0.68
Winter			473	0.70	507	0.63	434	0.58
Spring			452	0.65	506	0.51	409	0.53
District B: 2016-17								
Fall			140	0.58	162	0.64	157	0.58
Winter			141	0.70	158	0.62	151	0.68
Spring			126	0.66	163	0.64	151	0.57
Districts A & B: 2015-17								
Fall	373	0.65	1107	0.62	1183	0.64	1081	0.61
Winter	383	0.65	1110	0.68	1202	0.66	1062	0.60
Spring	381	0.69	1037	0.70	1192	0.62	1035	0.57

One natural question is whether these correlations differ for ELL students. Table 4 present results for ELL and non-ELL students respectively. To avoid small samples, and potentially noisy results, we only report findings for the pooled sample. While the results certainly differ between the two groups, there is no obvious pattern showing stronger correlations for one group over the other. Both groups generally show correlations in the .60-.70 range. These patterns suggest that Literably may be an appropriate screening tool for both ELL and non-ELL students.

The lowest correlations we observe in this table are found at the end of grade 5, for both ELL and non-ELL students, and the end of grade 4 for non-ELL students. This again hints at the possibility that WCPM scores are not effectively differentiating between students with relatively high reading abilities.

Table 4: Criterion Validity for Literably WCPM Score, Districts A and B, ELL and Non-ELL Students, 2015-2017

	Grade 2		Grade 3		Grade 4		Grade 5	
Season	N	Corr.	N	Corr.	N	Corr.	N	Corr.
ELL								
Fall	181	0.62	238	0.66	230	0.69	257	0.66
Winter	186	0.65	243	0.70	240	0.67	250	0.60
Spring	186	0.70	229	0.63	240	0.59	235	0.56
Non-ELL								
Fall	192	0.67	362	0.64	415	0.64	364	0.63
Winter	197	0.62	371	0.69	425	0.55	335	0.53
Spring	195	0.66	349	0.66	429	0.45	325	0.45

Clasification Accuracy

We turn now to classification accuracy. We report a variety of estimates intended to provide insight into how well the Literably assessment distinguished between students who will and will not meet the SBAC standards. Although we have focused exclusively on the SBAC ELA assessment to this point, here we focus on an additional SBAC outcome: meeting the SBAC Reading standard. For parsimonious presentation of the results, we provide estimates for only the pooled sample, although we continue to present results for ELL and non-ELL students.

Figures 2 (ELA) and 3 (Reading) report results from area under the receiver operating characteristic curve (ROC curve) analyses. Results for the full sample, which is the left-most panel in figures 2 and 3, show that all of the results fall into the *good* category (based on criteria discussed in Method section). These results are consistent with area under the ROC curve analyses conducted on previously researched and widely-used early literacy assessments (NCS Pearson 2012; Christ and Colleagues 2015).

The middle panels of figures 2 and 3 reports results for students classified as ELL. We generally observe that the ROC scores tend to be slightly lower for ELL students, especially when considering the Reading standard. Again, most of the results fall within the *good* category, but we also observe that, for the Reading results, a handful fall below this threshold.

Results for non-ELL students, presented in the right-most panel of figures 2 and 3, are, in general, slightly higher than those reported in the middle panel. In particular, the results for the Reading standard tend to be higher when compared to the ELL sample, although this is not true of grade 5. The majority of these results fall into the *good* category.

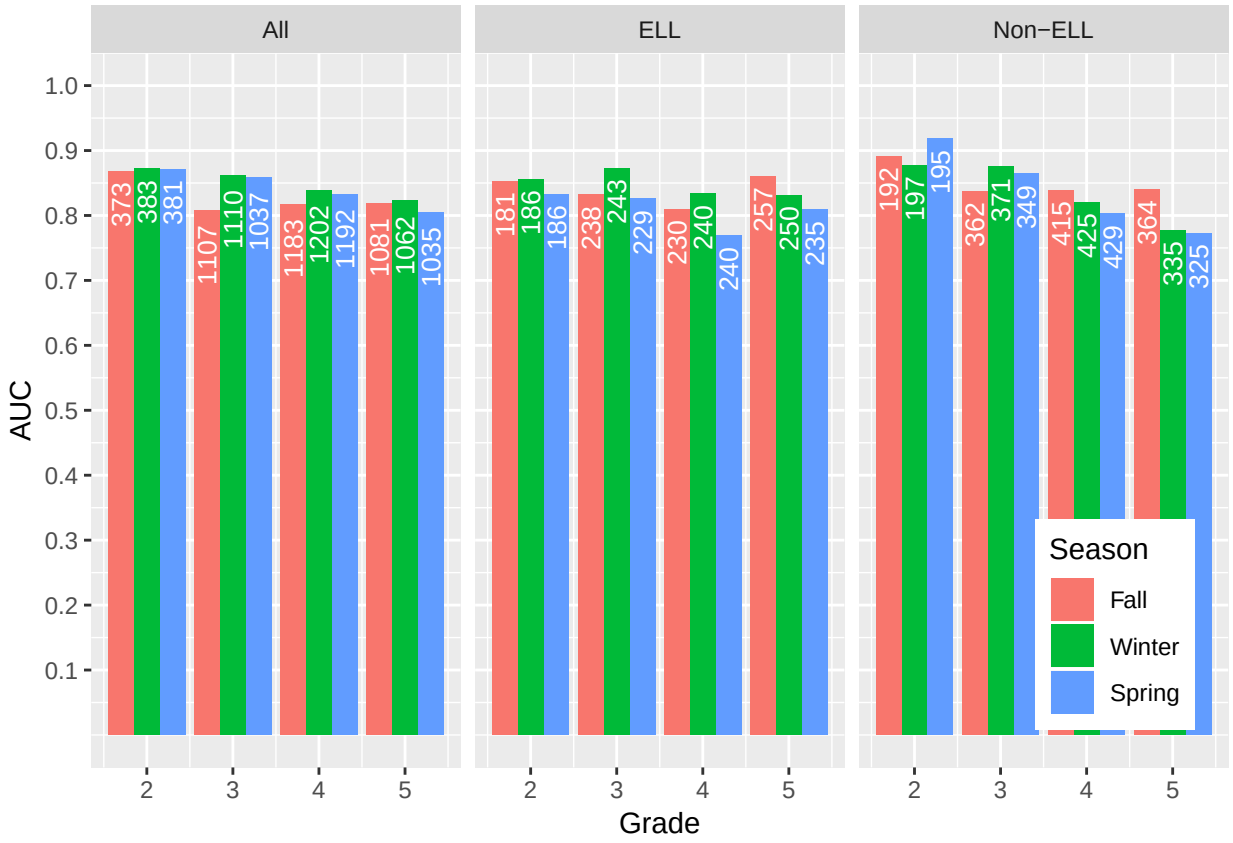

Figure 2: ELA: Area Under the ROC Curve, by Student ELL Status, Grade, and Time of Year (sample size shown inside each bar)

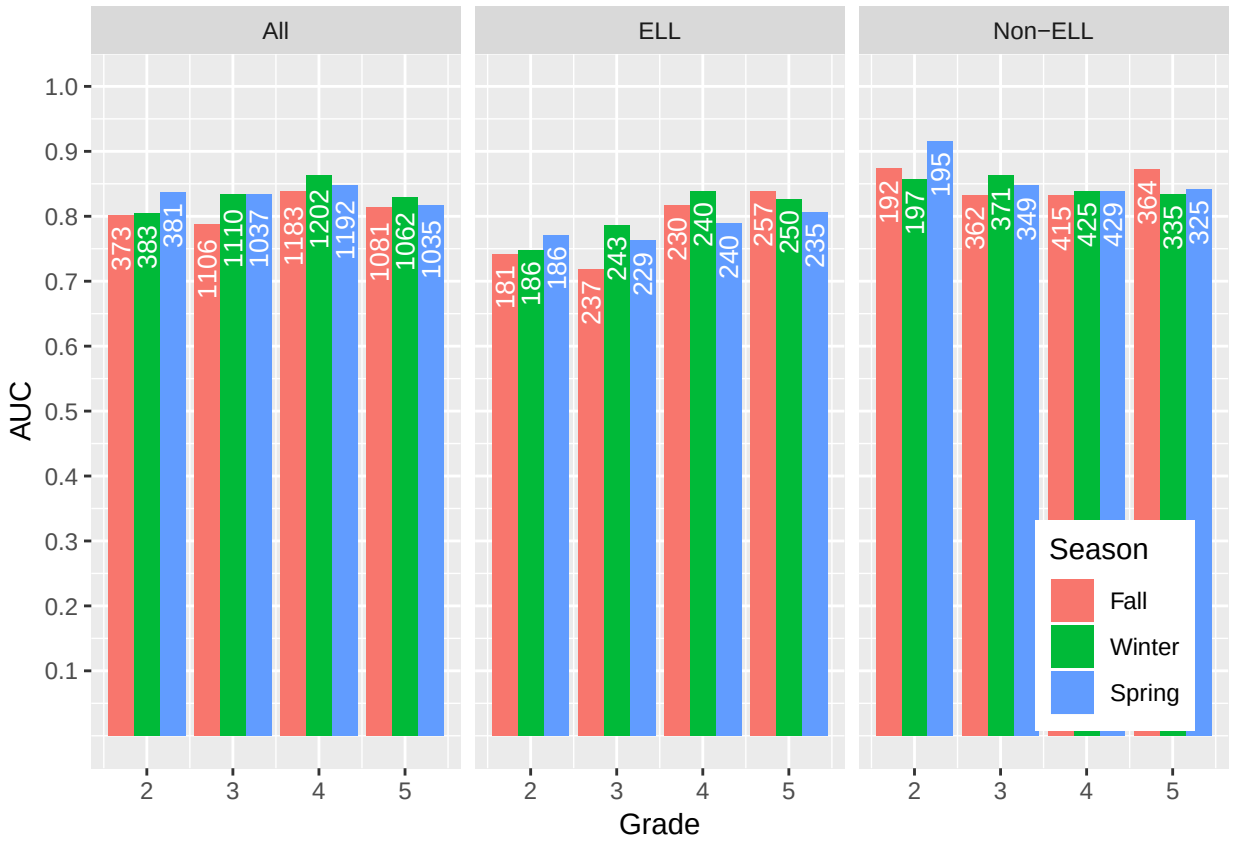

Figure 3: Reading: Area Under the ROC Curve, by Student ELL Status, Grade, and Time of Year (sample size shown inside each bar)

We also consider results based on logistic regressions, focusing specifically on Nagelkerke's R^2. Results for the ELA claim generally fall within, or near to the range of acceptable values as seen in the left-most panel of figures 4 and 5. The R^2 for meeting the ELA standard tend to be higher than the values for the Reading standard, although this is not true in grade 4. As a result, the Nagelkerke's R^2 values for the Reading claim are sometimes below the standard range. The highest R^2 values are found in grade 2, ELA. This again hints at the potential for survivor bias in the grade 2 results.

The middle and right panels of figures 4 and 5 reports Nagelkerke's R^2 values for ELL and non-ELL students. In these panels we observe R^2 values that are typically higher for meeting the ELA standard, as compared to the Reading standard. Comparing students classified as ELL to non-ELL students, we see no clear pattern in these results. In some cases the values are higher for ELL students, as we see in grade 5, while in other cases the results are stronger for non-ELL students, such as grade 2. The range of values across the panels and figures are similar, and, again, are in line with similar early literacy assessments.

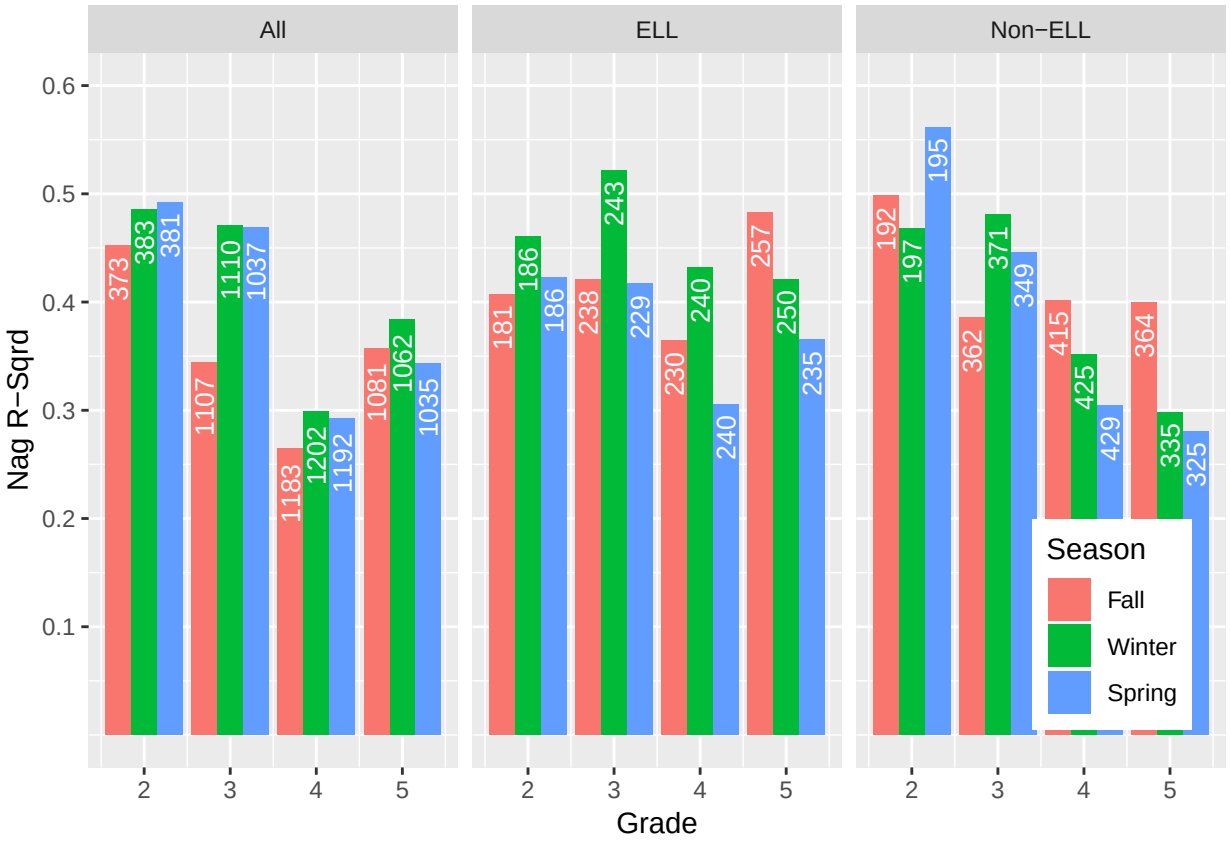

Figure 4: WCPM and SBAC ELA Logistic Regression: Nagelkerke's R-Sqrd, by Student ELL Status, Grade, and Time of Year (sample size shown inside each bar)

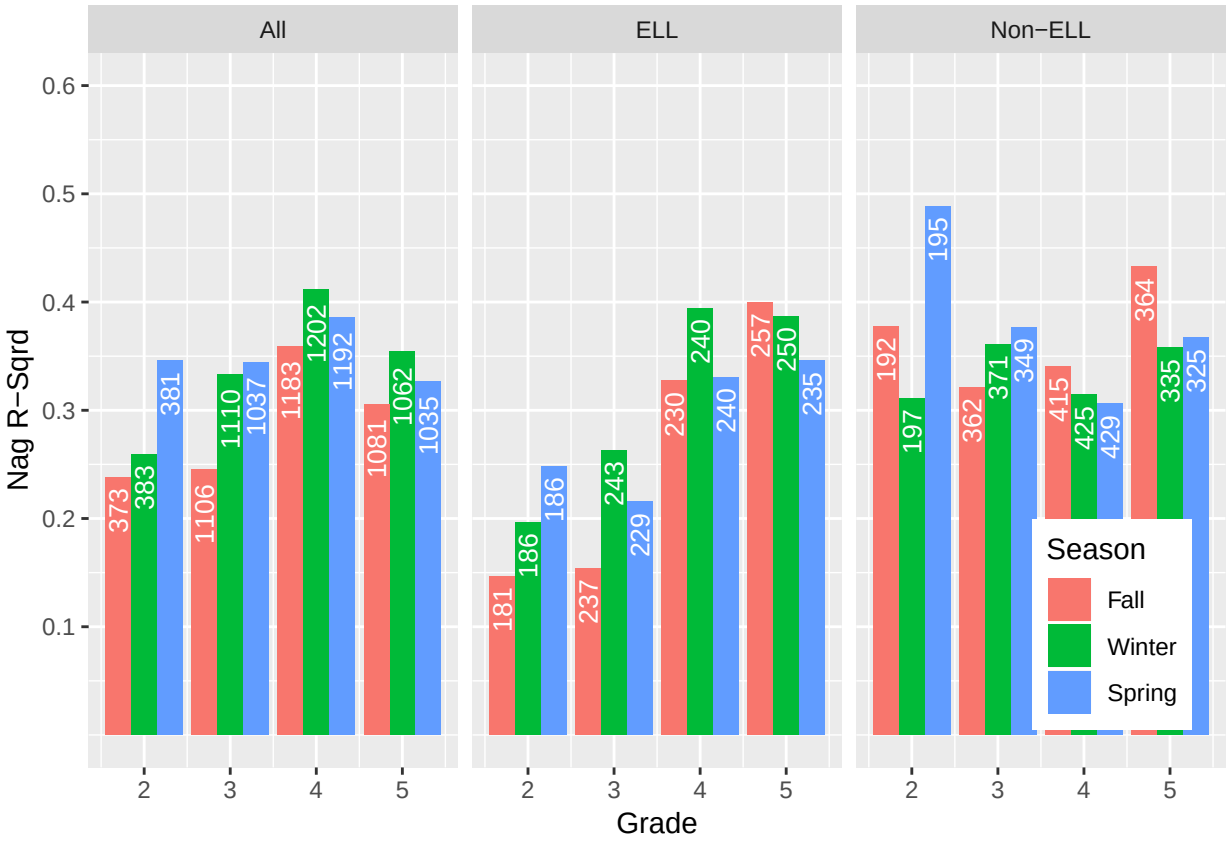

Figure 5: WCPM and SBAC Reading Logistic Regression: Nagelkerke's R-Sqrd, by Student ELL Status, Grade, and Time of Year (sample size shown inside each bar)

Cut Scores

The final evaluation of classification accuracy comes from examining the districts' use of the Literably-provided cut scores. Literably sets cut scores using data pooled from both districts (Townsend and Domingue 2018), so here we focus on how effectively the cut scores classify students within each district. Literably currently only provides cut scores for meeting the ELA standard, so no results are shown for meeting the Reading standard.

Tables 5 and 6 report results for District A, for the 2015-2016 and 2016-2017 academic years, respectively. The results are strong in both years for both sensitivity and specificity. These values tend to be in the .70-.80 range, and extend in some cases above .80. The percent correctly classified confirms that these cut scores are effective classifiers, as, on average, more than seven out of ten students are classified correctly. These results are similar to those reported by other publishers of early literacy assessments (NCS Pearson 2012; Christ and Colleagues 2015).

Table 5: District A, 2015-2016: Classification Performance of Literably Cut Scores

				Cut Score & Related Statistics			
Season	Prof. Level	Student Count	At-Level Count	Cut Score	Sens.	Spec.	Pct. Correct
Grade 2							
Fall	Met	373	255	51	0.71	0.83	0.79
Winter	Met	383	261	61	0.71	0.81	0.78
Spring	Met	381	260	71	0.74	0.83	0.80
Grade 3							
Fall	Met	507	329	70	0.53	0.87	0.75
Winter	Met	496	321	79	0.73	0.79	0.77
Spring	Met	459	286	82	0.81	0.79	0.80
Grade 4							
Fall	Met	538	348	86	0.60	0.82	0.74
Winter	Met	537	349	95	0.71	0.83	0.79
Spring	Met	523	342	108	0.82	0.78	0.79
Grade 5							
Fall	Met	460	341	95	0.63	0.72	0.70
Winter	Met	477	350	101	0.72	0.78	0.76
Spring	Met	475	341	102	0.65	0.80	0.76

Table 6: District A, 2016-2017: Classification Performance of Literably Cut Scores

Season	Prof. Level	Student Count	At-Level Count	Cut Score & Related Statistics			
				Cut Score	Sens.	Spec.	Pct. Correct
Grade 3							
Fall	Met	460	309	70	0.84	0.70	0.75
Winter	Met	473	317	79	0.89	0.70	0.77
Spring	Met	452	299	82	0.81	0.76	0.78
Grade 4							
Fall	Met	483	321	86	0.69	0.84	0.79
Winter	Met	507	342	95	0.72	0.81	0.78
Spring	Met	506	341	108	0.78	0.56	0.63
Grade 5							
Fall	Met	464	316	95	0.75	0.82	0.80
Winter	Met	434	284	101	0.80	0.68	0.72
Spring	Met	409	258	102	0.79	0.65	0.70

Table 7 reports results for District B in the 2016-2017 academic year. While results here are generally in the same range as those from District A, we do observe more variation in both the sensitivity and specificity values. For example, in some instances, specificity values are below .60, yet in other cases sensitivity values are greater than .90. It is not surprising, then, that the percent correctly classified results are slightly lower than in District A. Still, on average, nearly seven out of 10 students are classified correctly by the Literably-provided cut scores.

Table 7: District B, 2016-2017: Classification Performance of Literably Cut Scores

Season	Prof. Level	Student Count	At-Level Count	Cut Score & Related Statistics			
				Cut Score	Sens.	Spec.	Pct. Correct
Grade 3							
Fall	Met	140	85	70	0.91	0.42	0.61
Winter	Met	141	88	79	0.89	0.56	0.68
Spring	Met	126	79	82	0.81	0.68	0.73
Grade 4							
Fall	Met	162	115	86	0.72	0.70	0.70
Winter	Met	158	110	95	0.71	0.77	0.75
Spring	Met	163	113	108	0.88	0.65	0.72
Grade 5							
Fall	Met	157	93	95	0.78	0.63	0.69
Winter	Met	151	90	101	0.57	0.80	0.71
Spring	Met	151	90	102	0.69	0.78	0.74

Conclusion

This report presents the first results of a series of validity analyses on the Literably early literacy assessment. Focusing on criterion validity and classification accuracy, we find that in nearly all cases the properties of the Literably assessment are analogous to previously researched and widely-used early literacy screening tools such as those from DIBELS™, AIMSweb™, and easyCBM™. This is true when we analyze the data from the two districts separately as well as when we analyze a pooled dataset combining the two districts. While we see some differences in criterion validity and classification accuracy results for ELL and non-ELL students, even here the vast majority of reported results are commensurate with previously researched early literacy screening tools.

One minor concern in the criterion validity results was the lower observed correlations between WCPM scores and ELA SBAC scores for grades 4 and 5 in the spring trimester. The lower correlations observed here may be due to a ceiling effect in WCPM scores for students with relatively high reading abilities - recall that the mean WCPM scores for these students was roughly 110 in the winter and the spring. These results suggest that Literably might be able to make minor improvements to the assessment for grades 4 and 5, with an eye towards increased score differentiation for students with strong reading abilities.

With only two districts in the study, we cannot generalize these results to either other districts

in California or the country more broadly. It is worth noting, however, that the reported results are similar for the two districts, despite some clear differences between them such as location, size, and ELL composition. More research will be needed to know if the technical properties of the assessment are robust to districts with other types of differences.

Results presented here suggest that the Literably early literacy assessment performs similarly to other literacy screening tools such as those from DIBELS™, AIMSweb™, and easyCBM™. Given the core similarity of the measurement task, the similarity in results between Literably and its peers is perhaps to be expected. Literably, however, differs in a key way from its competitors and it is thus important to document that it has similar psychometric features.

References

Buck, Julie, and Joseph Torgesen. 2003. "The Relationship Between Performance on a Measure of Oral Reading Fluency and Performance on the Florida Comprehensive Assessment Test." *Tallahassee, FL: Florida Center for Reading Research.*

Christ, Theodore J., and Colleagues. 2015. "Formative Assessment System for Teachers: Abbreviated Technical Manual for Iowa Version 2.0." *Minneapolis, MN: Author and FastBridge Learning.*

Danne, Mary C, Jay R Campbell, Wendy S Grigg, Madeline J Goodman, and Andreas Oranje. 2005. "Fourth-Grade Students Reading Aloud: NAEP 2002 Special Study of Oral Reading. the Nation's Report Card. Nces 2006-469." *National Center for Education Statistics*. ERIC.

Fountas, Irene C, and Gay Su Pinnell. 1996. *Guided Reading: Good First Teaching for All Children.* ERIC.

Fuchs, Lynn S, Douglas Fuchs, Michelle K Hosp, and Joseph R Jenkins. 2001. "Oral Reading Fluency as an Indicator of Reading Competence: A Theoretical, Empirical, and Historical Analysis." *Scientific Studies of Reading* 5 (3). Taylor & Francis: 239–56.

Good, Roland H, RA Kaminski, EN Dewey, J Wallin, KA Powell-Smith, and RJ Latimer. 2013. "Dynamic Indicators of Basic Early Literacy Skills: DIBELS Next Technical Manual."

Good, Roland H, Kelly A Powell-Smith, Mary Abbott, Elizabeth N Dewey, Amy N Warnock, and Dave VanLoo. 2018. "Examining the Association Between DIBELS Next and the Sbac Ela Achievement Standard." *Contemporary School Psychology*. Springer, 1–12.

Good, Roland H, Deborah C Simmons, and Edward J Kame'enui. 2001. "The Importance and Decision-Making Utility of a Continuum of Fluency-Based Indicators of Foundational Reading Skills for Third-Grade High-Stakes Outcomes." *Scientific Studies of Reading* 5 (3). Taylor & Francis: 257–88.

Kilgus, Stephen P, Scott A Methe, Daniel M Maggin, and Jessica L Tomasula. 2014. "Curriculum-Based Measurement of Oral Reading (R-CBM): A Diagnostic Test Accuracy Meta-Analysis of Evidence Supporting Use in Universal Screening." *Journal of School Psychology* 52 (4). Elsevier: 377–405.

Literably. 2018. "Literably Techincal Manual." *San Francisco, CA: Literably, Inc.* https://s3.amazonaws.com/literably-assets/literably_technical_manual_latest.pdf.

NCS Pearson. 2012. "Aimsweb Technical Manual." *Bloomington, MN: Author. Pianta, R.,*

Howes, C., Burchinal, M., Bryant, D., Clifford, R., Early, D., & Barbarin, O.

Riedel, Brant W. 2007. "The Relation Between Dibels, Reading Comprehension, and Vocabulary in Urban First-Grade Students." *Reading Research Quarterly* 42 (4). Wiley Online Library: 546–67.

Salvia, John, James Ysseldyke, and Sara Witmer. 2012. *Assessment: In Special and Inclusive Education*. Cengage Learning.

Townsend, Joseph B., and Benjamin W. Domingue. 2018. "Setting Literably Cut Scores." *San Francisco, CA: Literably, Inc.* https://s3.amazonaws.com/literably-assets/literably_cut_scores.pdf.

Yang, Jiaxi, and Benjamin W. Domingue. 2018. "Setting Literably Cut Scores." *San Francisco, CA: Literably, Inc.* https://s3.amazonaws.com/literably-assets/literably_reliability_report.pdf.