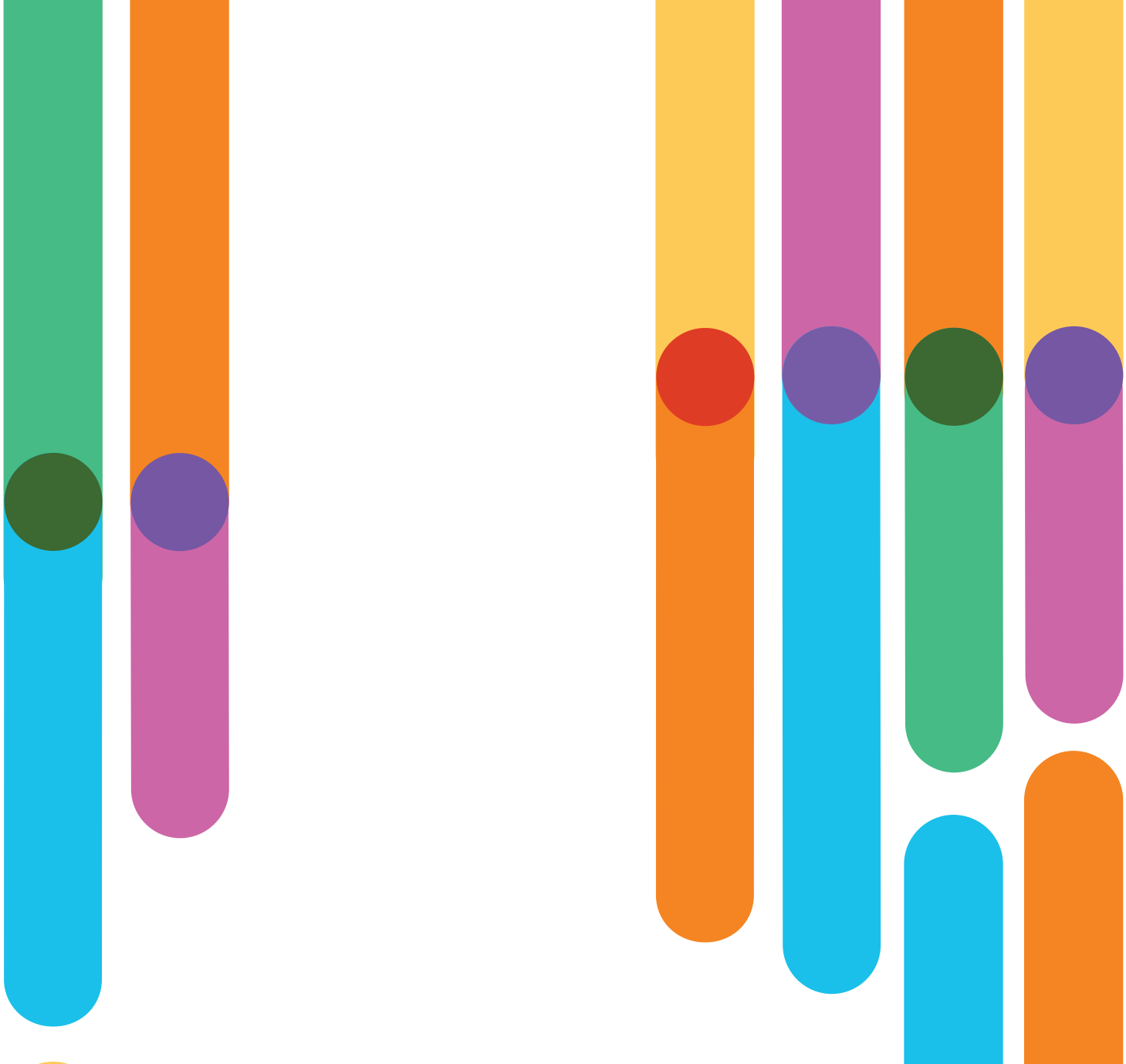

# Literably

## Technical Manual

Literably, Inc.
345 9th St
San Francisco, CA 94103
hello@literably.com
(415) 842-9802

# Table of Contents

# List of Tables & Figures

# Introduction

Literably is a literacy assessment for grades K-8 designed to inform instruction, screen students for reading difficulties, identify students' reading levels, and monitor students' response to intervention. The data needed to do this are generated by a student reading a short passage out loud into a device and answering several comprehension questions.

In this manual, we intend to provide educators and administrators with information that helps them decide whether to use Literably, and, for Literably's users, how to use Literably appropriately. We report technical findings supporting the adequacy of Literably for universal screening, as well as background information about the design of the assessment and the research that guided its construction.

Our conclusions parallel those of decades of research on what teachers can learn about their readers by listening to them read out loud. Ever since Deno et. al (1980) demonstrated the remarkable correlation between a student's speed at reading a passage and her achievement on standardized reading tests, a large body of research has demonstrated that oral reading fluency is a particularly effective tool for identifying struggling readers and measuring their progress. This effectiveness is confirmed in the technical documentation accompanying widely used oral reading assessments, such as the Dynamic Indicators of Basic Early Literacy Skills (Good & Kaminski 2002) or AIMSweb (Howe & Shinn 2002).

Like these well-known instruments, Literably yields an oral reading fluency (ORF) score. An ORF score measures the number of words a student reads out loud correctly in one minute and is sometimes written as *words correct per minute*. ORF has the desirable property that, in general, it is valid for the purpose of screening elementary school students for reading difficulties (Kilgus, Methe, Maggin, & Tomasula 2014). Publishers of screening tools typically present evidence showing the strength of the statistical relationships between a student's ORF score and her performance on high-stakes reading assessments, either concurrently or at a later date. If strong, these relationships—along with other relevant information—allow an assessment's results to be used with confidence to classify students according to their risk of reading failure and identify students in need of intervention.

Unlike students assessed under these older instruments, students that use Literably read out loud into a device, typically a tablet or laptop. This is in contrast to the traditional process of an educator sitting with a student one-on-one and noting her errors as she reads. This change in mode of administration saves a large amount of staff time, which can improve student achievement via the reallocation of that time toward instructional activities (Adams 2005). But because student behavior, scoring procedures, level of psychological support, student comfort, and many other variables could differ depending on administration mode, the validity of the assessment might also differ in the shift to online administration.

To answer this question, researchers examined the correspondence between

Literably ORF scores and scores on the Smarter Balanced Assessment Consortium (SBAC) for students in grades 2-5 in two very different school districts. In both settings, Literably scores show both a strong overall correlation with SBAC ELA scores (about .65) and a high level of accuracy in identifying students with a low SBAC achievement level (about 75%). These values are close to the values reported for other widely used teacher-administered assessments, such as DIBELS (Good et. al. 2017) or AIMSWeb (NCS Pearson 2012). Taken together with descriptive information on the content and structure of the assessment, these results support the use of Literably for universal screening in reading in the elementary grades.

# Development

## Passages

To promote content validity, Literably's passages were selected to resemble the materials that students read in the classroom. Thus, Literably draws its reading passages, with permission, from leveled children's trade books. Approximately half are excerpts and the remainder are the full book, with the full books concentrated at the lowest reading levels. Approximately half are fiction and half non-fiction. Pictures are available for all passages up through the 3rd grade level. All books were officially leveled according to Heinemann's Guided Reading framework (Fountas & Pinnell, 1996).

A team of five former or current elementary educators selected the texts and excerpts from the publishers' catalogs. All had experience either as a teacher of relevant grades or as a reading specialist. A Literably employee then examined each proposed passage and reviewed it along the dimensions of age-appropriateness, background knowledge required, and syntactic and phonic difficulty. Several school and district administrators experienced with assessment selection assisted with the process. When necessary, Literably staff made small adjustments to texts to improve their conformity along these dimensions.

The following distribution of passages at each guided reading level emerged from the passage development process:

**Table 1:** Passage Distribution by Reading Level and Grade

| | | | |
|---|---|---|---|
| A (K) | 3 | N (3rd) | 2 |
| B (K) | 2 | O (3rd) | 2 |
| C (K) | 2 | P (3rd) | 2 |
| D (K) | 3 | Q (4th) | 2 |
| E (1st) | 3 | R (4th) | 2 |
| F (1st) | 3 | S (4th) | 2 |
| G (1st) | 2 | T (5th) | 2 |
| H (1st) | 2 | U (5th) | 2 |
| I (1st) | 2 | V (5th) | 2 |
| J (1st) | 2 | W (5th) | 2 |
| K (2nd) | 2 | X (6th) | 2 |
| L (2nd) | 2 | Y (6th) | 2 |
| M (2nd) | 2 | Z (7th) | 2 |

## Equating

Since the levels of Literably's passages are spread out across Guided Reading levels, they vary in difficulty markedly, even across passages of the same grade level. A student's reading fluency rate thus would be expected to vary across alternate forms, diminishing the technical adequacy of the assessment. Even under assessments with passage sets deliberately designed to be equivalent, the passage administered has a substantial impact on a student's score (Betts, Pickart, & Heistad, 2009).

To minimize differences due to these alternate forms, the ORF scores for most passages in Literably's corpus were statistically equated. To do this, Literably staff selected an anchor passage for each grade level K-7. The anchor for each grade is the passage in the Literably database with the nearest ORF scores to national oral reading fluency norms for that grade (Hasbrouck & Tindal, 2017). Students' relative performance levels on different Literably passages were then used to develop an equating table linking all other passages to each anchor passage. Where data is insufficient to compute an equated score, the non-equated score is reported.

When reporting scores to end users, Literably reports both the raw reading rate and the equated reading rate. For most purposes, including universal screening, we recommend using the equated scores. Literably's norms and cut scores for risk (available in the Appendix) were drawn from the equated scores.

## Raters

The oral reading fluency scores are computed from the transcriptions of Literably's raters. Literably's raters are selected by a competitive application process wherein successful applicants must transcribe a sample of student audio recordings with a high rate of concordance with the transcriptions of Literably staff members. Once hired, raters regularly face random test recordings to ensure that their performance levels remain high; low performers are dismissed.

## Comprehension Questions

The authoring of the comprehension items followed a similar process to the development of the reading passages. Literably staff worked with a team of experienced educators to write the items and then revised them over time with feedback from users.

To maximize alignment with the curricula of schools in the United States, the authors wrote every item to assess one or more of the Common Core State Standards (Common Core State Standards Initiative, 2010). Because Literably's passages may be read by students of any grade level, staff aligned the items to the K-5 anchor standards rather than to the grade-level standards.

## Cut Scores

Literably compares students' ORF scores to cut scores for risk based on grade and season. For grades 2-5, the scores were derived empirically from student performance on the Smarter Balanced Assessment Consortium (SBAC) in two California school districts (Townsend & Domingue 2018). The goal of the cut score setting process was to maximize overall accuracy at predicting reading success or failure, where success is defined as Achievement Level 3 or 4 on the SBAC.

On this basis, we later report classification accuracy for grades 2-5. We expect adjustments to both the cut scores and their performance as additional data arrive and the cut score process is further improved.

The current cut scores themselves can be viewed in the Appendix. Note that the scores occasionally decline from the spring of one grade to the fall of the following, perhaps reflecting a combination of summer learning loss and the jump in difficulty of grade-level text.

# Validity

## Correlations

Table 2 reports the correlations of Literably ORF scores and ELA scores on the SBAC. The data are drawn from two California school districts during the 2015-2016 and 2016-2017 school years, one a midsize urban district and the other a small rural district (Townsend & Domingue 2018). The correlations are about .65, on par with the Dynamic Indicators of Basic Early Literacy Skills (DIBELS) and AIMSWeb (Good et. al. 2017; NCS Pearson 2012).

Some of these correlations are slightly lower for technical reasons than they would be otherwise. Literably scores are not available in equated form prior to the winter season of the 2015-2016 school year. As a result, the fall correlations would be somewhat higher if the assessments were administered today. Additionally, since the SBAC is not administered until the third grade, the correlation for grade 2 reflects a one-year longer lag than the others and thus would be expected to be somewhat lower than it would be with a criterion administered closer in time.

**Table 2:** Criterion Validity of the Literably ORF Score

| Grade (Season) | *N* | Correlation |
|---|---|---|
| 2 (Fall) | 373 | .65 |
| 2 (Winter) | 383 | .65 |
| 2 (Spring) | 381 | .69 |
| 3 (Fall) | 1107 | .62 |
| 3 (Winter) | 1110 | .68 |
| 3 (Spring) | 1037 | .70 |
| 4 (Fall) | 1183 | .64 |
| 4 (Winter) | 1202 | .66 |
| 4 (Spring) | 1192 | .62 |
| 5 (Fall) | 1081 | .61 |
| 5 (Winter | 1062 | .60 |
| 5 (Spring) | 1035 | .57 |

Tables 3 and 4 contain similar correlations for the two California school districts separately, indicating the generalizability of the relationship demonstrated in Table 2 across multiple settings. Table 3 shows the correlation between Literably and the SBAC by year of administration for District A, a midsize urban district. There is considerably more variation in the values of the coefficients than when the samples were pooled in Table 2, which may be partly due to the smaller sample sizes.

**Table 3:** Criterion Validity of the Literably ORF Score, District A

| Grade (Season) | Year | *N* | Correlation |
|---|---|---|---|
| 2 (Fall) | 2017 | 373 | .65 |
| 2 (Winter) | 2017 | 383 | .65 |
| 2 (Spring) | 2017 | 381 | .69 |
| 3 (Fall) | 2016 | 507 | .63 |
| 3 (Winter) | 2016 | 496 | .67 |
| 3 (Spring) | 2016 | 459 | .74 |
| 3 (Fall) | 2017 | 460 | .66 |
| 3 (Winter) | 2017 | 473 | .70 |
| 3 (Spring) | 2017 | 452 | .74 |
| 4 (Fall) | 2016 | 538 | .60 |
| 4 (Winter) | 2016 | 537 | .70 |
| 4 (Spring) | 2016 | 523 | .73 |
| 4 (Fall) | 2017 | 483 | .70 |
| 4 (Winter) | 2017 | 507 | .63 |
| 4 (Spring) | 2017 | 506 | .51 |
| 5 (Fall) | 2016 | 460 | .52 |
| 5 (Winter | 2016 | 477 | .61 |
| 5 (Spring) | 2016 | 475 | .60 |
| 5 (Fall) | 2017 | 464 | .68 |
| 5 (Winter) | 2017 | 434 | .58 |
| 5 (Spring) | 2017 | 409 | .53 |

Table 4 shows the same analysis for District B, a small rural district. All the correlations in Table 4 are between Literably ORF scores during the 2016-2017 school year and the spring 2017 SBAC.

**Table 4:** Criterion Validity of the Literably ORF Score, District B

| Grade (Season) | *N* | Correlation |
|---|---|---|
| 3 (Fall) | 140 | .58 |
| 3 (Winter) | 141 | .70 |
| 3 (Spring) | 126 | .66 |
| 4 (Fall) | 162 | .64 |
| 4 (Winter) | 158 | .62 |
| 4 (Spring) | 163 | .64 |
| 5 (Fall) | 157 | .58 |
| 5 (Winter | 151 | .68 |
| 5 (Spring) | 151 | .57 |

Because of the potential advantages of individual administration for interpreting the spoken language of English Language Learners (ELLs), questions naturally arise about Literably's validity for this population. To resolve these questions, Table 5 reports the criterion validity of the ORF score for students in District A for the 2017 SBAC, revealing little difference:

**Table 5:** Criterion Validity of the Literably ORF Score for ELL Students

| Grade (Season) | *N* | ELL Corr. | *N* | Non-ELL Corr. |
|---|---|---|---|---|
| 2 (Fall) | 181 | .62 | 192 | .67 |
| 2 (Winter) | 186 | .65 | 197 | .62 |
| 2 (Spring) | 186 | .70 | 195 | .66 |
| 3 (Fall) | 225 | .64 | 235 | .65 |
| 3 (Winter) | 231 | .70 | 242 | .66 |
| 3 (Spring) | 218 | .64 | 234 | .64 |
| 4 (Fall) | 228 | .69 | 255 | .64 |
| 4 (Winter) | 238 | .66 | 269 | .52 |
| 4 (Spring) | 238 | .59 | 268 | .36 |
| 5 (Fall) | 250 | .65 | 214 | .63 |
| 5 (Winter | 245 | .59 | 189 | .49 |
| 5 (Spring) | 229 | .55 | 180 | .42 |

## Logistic Regression

Logistic regression is a statistical procedure for predicting a binary variable from independent variable(s) of interest. Whereas the correlations previously reported depend upon the association between the Literably ORF score and the SBAC score across the entire spectrum of scores, logistic regression uses the Literably ORF score to predict the binary question of whether a student achieves reading success or failure on the SBAC. Nagelkerke's R2 is a commonly used measure of how closely this prediction fits the underlying data (ranging from 0 to 1). Table 6 reports the results of a logistic regression of 2017 SBAC ELA scores on Literably ORF scores for District A. These results compare favorably to similar assessments, such as DIBELS (Good et. al. 2017).

**Table 6:** Logistic Regression of SBAC Scores on the Literably ORF Scores

| Grade (Season) | *N* | Nagelkerke's $R^2$ |
|---|---|---|
| 2 (Fall) | 373 | .45 |
| 2 (Winter) | 383 | .49 |
| 2 (Spring) | 381 | .49 |
| 3 (Fall) | 460 | .45 |
| 3 (Winter) | 473 | .54 |
| 3 (Spring) | 452 | .47 |
| 4 (Fall) | 483 | .44 |
| 4 (Winter) | 507 | .42 |
| 4 (Spring) | 506 | .28 |
| 5 (Fall) | 464 | .50 |
| 5 (Winter | 434 | .40 |
| 5 (Spring) | 409 | .35 |

# Classification Accuracy

Classification accuracy is a general term for the effectiveness of a screening instrument in predicting a binary criterion. Three common measures for doing this, which we report here for Literably, are sensitivity, specificity, and the area under the ROC curve:

1. *Sensitivity* is the proportion of students that failed the criterion that were predicted to do so by the instrument and a particular cut score.
2. *Specificity* is the proportion of students predicted to succeed by the instrument instrument and a particular cut score that actually succeeded.
3. *Area under the ROC curve* is a measure that evaluates an instrument's overall potential for sensitivity and specificity across a range of possible cut scores.

For ORF assessments, the binary criterion is typically success or failure on a standardized end-of-grade reading assessment. In Table 7, we report these measures for Literably using Literably's cut scores for reading success and student performance on a student's next Smarter Balanced Assessment Consortium (SBAC) in the two California school districts referenced earlier (Townsend & Domingue, 2018).

**Table 7:** Classification Accuracy of Literably ORF, by Grade and Season

| Grade (Season) | *N* | AUC | Sensitivity | Specificity |
|---|---|---|---|---|
| 2 (Fall) | 373 | .87 | .71 | .83 |
| 2 (Winter) | 383 | .87 | .75 | .81 |
| 2 (Spring) | 381 | .87 | .70 | .85 |
| 3 (Fall) | 1107 | .81 | .71 | .74 |
| 3 (Winter) | 1110 | .86 | .71 | .82 |
| 3 (Spring) | 1037 | .86 | .71 | .86 |
| 4 (Fall) | 1183 | .82 | .71 | .79 |
| 4 (Winter) | 1202 | .84 | .70 | .83 |
| 4 (Spring) | 1192 | .83 | .71 | .78 |
| 5 (Fall) | 1081 | .82 | .70 | .76 |
| 5 (Winter) | 1062 | .82 | .70 | .78 |
| 5 (Spring) | 1035 | .80 | .72 | .74 |

Because the cut scores used to compute the classification accuracy results were generated using the data from these two districts, questions might arise about their generalizability. To address these questions, in Tables 8 and 9, we also report these figures separately for District A and District B.

These two districts are particularly different in their proportions of ELLs. Approximately 49% of District A's students are ELLs, while only about 7% of District B's are ELLs. The optimal cut scores on ORF assessments tend to be significantly different for students still learning English from fluent English speakers (Johnson et. al. 2009). The disaggregated classification accuracy is thus particularly informative for Literably's generalizability. (All data in Table 9 is drawn from criterion year 2017.)

**Table 8:** Classification Accuracy of Literably ORF, District A

| Grade (Season) | Year | *N* | AUC | Sensitivity | Specificity |
|---|---|---|---|---|---|
| 2 (Fall) | 2017 | 373 | .87 | .71 | .83 |
| 2 (Winter) | 2017 | 383 | .87 | .75 | .79 |
| 2 (Spring) | 2017 | 381 | .87 | .70 | .77 |
| 3 (Fall) | 2016 | 507 | .82 | .53 | .87 |
| 3 (Winter) | 2016 | 496 | .85 | .63 | .86 |
| 3 (Spring) | 2016 | 459 | .87 | .73 | .87 |
| 3 (Fall) | 2017 | 460 | .86 | .84 | .70 |
| 3 (Winter) | 2017 | 473 | .89 | .75 | .81 |
| 3 (Spring) | 2017 | 452 | .86 | .67 | .86 |
| 4 (Fall) | 2016 | 538 | .79 | .65 | .80 |
| 4 (Winter) | 2016 | 537 | .84 | .70 | .86 |
| 4 (Spring) | 2016 | 523 | .88 | .70 | .88 |
| 4 (Fall) | 2017 | 483 | .85 | .75 | .82 |
| 4 (Winter) | 2017 | 507 | .84 | .70 | .82 |
| 4 (Spring) | 2017 | 506 | .77 | .69 | .70 |
| 5 (Fall) | 2016 | 460 | .76 | .62 | .74 |
| 5 (Winter) | 2016 | 477 | .83 | .69 | .80 |
| 5 (Spring) | 2016 | 475 | .80 | .65 | .80 |
| 5 (Fall) | 2017 | 464 | .87 | .74 | .83 |
| 5 (Winter) | 2017 | 434 | .83 | .78 | .72 |
| 5 (Spring) | 2017 | 409 | .80 | .79 | .65 |

| **Table 9:** Classification Accuracy of Literably ORF, District B | | | | |
|---|---|---|---|---|
| **Grade (Season)** | ***N*** | **AUC** | **Sensitivity** | **Specificity** |
| 3 (Fall) | 140 | .79 | .91 | .42 |
| 3 (Winter) | 141 | .85 | .81 | .68 |
| 3 (Spring) | 126 | .82 | .72 | .80 |
| 4 (Fall) | 162 | .82 | .77 | .66 |
| 4 (Winter) | 158 | .83 | .71 | .80 |
| 4 (Spring) | 163 | .87 | .82 | .75 |
| 5 (Fall) | 157 | .79 | .75 | .63 |
| 5 (Winter) | 151 | .80 | .54 | .84 |
| 5 (Spring) | 151 | .80 | .69 | .78 |

Overall, the values are similar to those reported by individually reported ORF assessments generally regarded as valid. AIMSWeb, for example, reports area under the curve values for R-CBM probe scores from grades 2-5 in the mid-to-high .8s (NCS Pearson 2012); Literably's are approximately .85. Similarly, Literably's sensitivity and specificity cluster around the .7s, much in line with DIBELS (Dewey et. al. 2013) and AIMSWeb (NCS Pearson 2012). These findings support Literably's usage to identify students in need of intervention to succeed as readers on the SBAC and similar high-stakes assessments. (They do not, however, provide evidence of Literably's ability to reliably monitor a student's response to intervention. Results on this question are forthcoming.)

# References


Adams, M. J. (2005). The promise of automatic speech recognition for fostering literacy growth in children and adults. *Handbook of literacy and technology*, 2, 109-128.

Betts, J., Pickart, M., & Heistad, D. (2009). An investigation of the psychometric evidence of CBM-R passage equivalence: Utility of readability statistics and equating for alternate forms. *Journal of School Psychology*, 47 (1), 1-17.

Common Core State Standards Initiative (2010). *Common Core State Standards for English Language Arts & Literacy in History/Social Studies, Science, and Technical Subjects*. Available: http://www.corestandards.org/wp-content/uploads/ELA_Standards1.pdf

Cummings, K. D., Biancarosa, G., Schaper, A., & Reed, D. K. (2014). Examiner error in curriculum-based measurement of oral reading. *Journal of School Psychology*, 52 (4), 361-375.

Deno, S. L., Mirkin, P. K., Chiang, B., and Lowry, L. (1980). *Relationships Among Simple Measures of Reading and Performance on Standardized Achievement Tests*. (Research Report No. 20). Minneapolis, MN: Institute for Research on Learning Disabilities, University of Minnesota.

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

Dewey, E. N., Powell-Smith, K. A., Good, R. H., & Kaminski, R. A. (2015). *DIBELS Next Technical Adequacy Brief*. Eugene, OR: Dynamic Measurement Group, Inc.

Fountas, I. C., & Pinnell, G. S. (1996). *Guided reading: Good first teaching for all children*. Portsmouth, NH: Heinemann.

Good, R. H., & Kaminski, R. A. (Eds.) (2002). *Dynamic Indicators of Basic Early Literacy Skills (6th ed.)*. Eugene, OR: Institute for the Development of Educational Achievement. Available: http://dibels.uoregon.edu.

Good, R. H., Powell-Smith, K. A., Abbott, M., Dewey, E. N., Warnock, A. N., & VanLoo, D. (2017). *Examining the Association Between DIBELS Next® and the SBAC ELA Achievement Standard*. Eugene, OR: Dynamic Measurement Group.

Available: https://dibels.org/Handout_Keynote_by_Roland_Good_2017_DIBELS_Super_Institute.pdf.

Hasbrouck, J., & Tindal, G. (2017). *An Update to Compiled ORF Norms. (Technical Report No. 1702)*. Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Howe, K. B., & Shinn, M. M. (2002). *Standard Reading Assessment Passages (RAPs) for use in General Outcome Measurement: A manual describing development and technical features*. Eden Prairie, MN: Edformation, Inc.

Johnson, E. S., Jenkins, J. R., Petscher, Y., & Catts, H. W. (2009). How can we improve the accuracy of screening instruments? *Learning Disabilities Research & Practice*, 24 (4), 174-185.

Kilgus, S. P., Methe, S. A., Maggin, D. M., & Tomasula, J. L. (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), 377-405.

National Reading Panel. (2000). *Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction: Reports of the subgroups*. Bethesda, MD: National Institute of Child Health and Human Development. Available: http://www.nationalreadingpanel.org/.

NCS Pearson. (2012). *AIMSWeb Technical Manual*. Bloomington, MN: NCS Pearson.

Townsend, J. & Domingue, B. (2018). *Predictive and Concurrent Validity of Literably Oral Reading Fluency (Technical Report No. 1)*. San Francisco, CA: Literably, Inc. Available: https://s3.amazonaws.com/literably-assets/literably_validity_report_1.pdf.

# Appendix: Cut Scores

| Grade | Fall | Winter | Spring |
|---|---|---|---|
| 2 | 51 | 62 | 69 |
| 3 | 70 | 71 | 73 |
| 4 | 89 | 93 | 100 |
| 5 | 94 | 97 | 102 |