ets research institute

AUGUST 2024
RM–24-06

RESEARCH MEMORANDUM

Replicating Decoding Threshold in ReadBasix®: Impact on Reading Skills Development

AUTHORS

Zuowei Wang, Tenaha O'Reilly, and Rebecca Sutherland

ETS Research Memorandum Series

Since its 1947 founding, ETS has conducted and disseminated scientific research to support its products and services, and to advance the measurement and education fields. In keeping with these goals, ETS is committed to making its research freely available to the professional community and to the general public. Published accounts of ETS research, including papers in the ETS Research Memorandum series, undergo a formal peer-review process by ETS staff to ensure that they meet established scientific and professional standards. All such ETS-conducted peer reviews are in addition to any reviews that outside organizations may provide as part of their own publication processes. Peer review notwithstanding, the positions expressed in the ETS Research Memorandum series and other published accounts of ETS research are those of the authors and not necessarily those of the Officers and Trustees of Educational Testing Service.

The Daniel Eignor Editorship is named in honor of Dr. Daniel R. Eignor, who from 2001 until 2011 served the Research and Development division as Editor for the ETS Research Report series. The Eignor Editorship has been created to recognize the pivotal leadership role that Dr. Eignor played in the research publication process at ETS.

Replicating Decoding Threshold in ReadBasix®: Impact on Reading Skills Development

Zuowei Wang and Tenaha O'Reilly
ETS Research Institute, Princeton, New Jersey, United States

Rebecca Sutherland
AERDF, Oakland, California, United States

August 2024

Corresponding author: Zuowei Wang, E-mail: zwang@ets.org

Suggested citation: Wang, Z., O'Reilly, T., & Sutherland, R. (2024). *Replicating decoding threshold in ReadBasix®: Impact on reading skills development* (Research Memorandum No. RM-24-06). ETS.

Find other ETS-published reports by searching the
ETS ReSEARCHER database.

To obtain a copy of an ETS research report, please visit
https://www.ets.org/contact/additional/research.html

Action Editor: Jamie Mikeska

Reviewers: Yang Jiang and Michael Flor

Abstract

In 2022, Advanced Education Research and Development Fund's (AERDF) Reading Reimagined launched the Equitable Foundational Literacy Research (EFLR) cohort, a group of four independently designed and researched instructional pilots aiming to support the foundational literacy skills of students from traditionally underserved populations in Grades 3–8. As part of the outcome metrics projects, EFLR used ReadBasix®, an ETS-developed reading component skills assessment, as a common measure. To help inform the use of ReadBasix and the interpretation of its scores, AERDF contracted ETS to conduct new analyses to replicate previously published decoding threshold studies with ReadBasix data. This research memorandum summarizes the findings of two replication studies. Study 1 first identified the location of the decoding threshold on the ReadBasix scoring scale (Word Recognition and Decoding score = 225) and then demonstrated the impact of scoring below the decoding threshold on the growth trajectories of other reading component skills measured by ReadBasix, including vocabulary, morphology, sentence processing, basic reading efficiency, and reading comprehension. Students scoring below the decoding threshold had significantly slower growth rate in all these skills. Study 2 replicated the behavioral differences between students scoring above versus below the decoding threshold when they responded to different kinds of ReadBasix decoding items. Students with poor decoding skills spent less time than peers when encountering a word that they did not yet know, indicating a lack of phonological recoding effort. Collectively, these results replicated earlier published studies with ReadBasix data, thus providing more evidence supporting the validity of ReadBasix and the robustness of the decoding threshold findings.

Keywords: ReadBasix®, decoding threshold, comprehension, literacy skills development

Acknowledgments

Rebecca Sutherland is the associate director, research for Reading Reimagined, a program of AERDF. AERDF is a national nonprofit dedicated to advancing research and development in PreK-12 education. Founded in 2021, AERDF pursues positive, multigenerational change by unlocking scientific discoveries and creating innovative solutions that improve teaching, learning, and assessment systems within education. Please visit https://aerdf.org/programs/readingreimagined for information on Reading Reimagined. The opinions expressed are those of the authors and do not necessarily represent views of Reading Reimagined or AERDF.

In the science of reading, decoding is the process a reader engages in to recognize words by applying knowledge in letter–sound correspondence. All major reading theories recognize the importance of decoding. According to the *Simple View of Reading* (Gough & Tunmer, 1986; Hoover & Gough, 1990), reading comprehension is the product of decoding and linguistic comprehension. That is, the reader applies decoding skills to sound out and recognize words (i.e., get words off the page), and successful decoding allows the reader to translate the reading task into a (spoken) language task.

In the lexical quality hypothesis (Perfetti & Hart, 2002), decoding is part of the orthographic system that enables the reader to recognize words and access their semantic information. Inefficient word recognition "would jeopardize comprehension processes that depend on a high quality representation" (Perfetti & Hart, 2002, p. 190). It follows that if decoding and word recognition skills fall below a threshold, comprehension becomes virtually impossible. Indeed, Wang et al. (2019) identified a *decoding threshold*: when a student's decoding score was below 235 on the RISE Word Recognition and Decoding subtest (WRD) (Sabatini et al., 2019), their performance on the RISE Reading Comprehension (RC) subtest was uniformly low.

The self-teaching hypothesis (Share, 1995) identifies decoding as a key driving factor for reading acquisition. Decoding enables the developing reader to translate a print word that is not yet readily recognizable into spoken language: If the reader recognizes the word from spoken language, the decoding process will have provided an opportunity for the reader to establish links between how the word is spelled (orthography), how the word is pronounced (phonology), and what the word means (semantics). These processes are some of the hallmarks of word learning. If the reader does not recognize the word from spoken language, the phonological recoding process still helps establish partial links among the three components, which facilitates future learning of the word. A reasonable deduction of the self-teaching hypothesis is that the self-teaching mechanism requires a minimum level of decoding skills. When decoding skills are inadequate, reading development stalls, which can be captured by slower growth in other reading subskills. Wang et al. (2019) discovered that students who scored below the decoding threshold on the WRD subtest of the RISE assessment

demonstrated limited reading comprehension growth in subsequent years. Furthermore, poor decoding skills and stagnant reading growth may impair the developing reader's motivation to read more (O'Reilly et al., 2019) and less reading in turn contributes to poor reading comprehension development (Mol & Bus, 2011).

Referred to as *phonics* and *word recognition*, decoding is one of four foundational skills that are necessary and important components of text comprehension in the Common Core State Standards (National Governors Association Center for Best Practices & Council of Chief State School Officers, 2010). Recognizing the importance of foundational reading skills, in 2022 the Reading Reimagined team at the Advanced Education Research and Development Fund (AERDF), a U.S. nonprofit organization established in 2021, funded instructional pilot research projects across the United States aimed at improving word recognition and fluency of traditionally underserved students from Grades 3–8. These projects are called the Equitable Foundational Literacy Research (EFLR) cohort[1] and requested to use ETS's ReadBasix®, a foundational reading skills battery (formerly called RISE) as a common outcome measure in their studies. The decoding threshold findings (Wang et al., 2019, 2020) have direct implications for intervention projects using ReadBasix as an outcome measure. Because the relationship between decoding and reading comprehension (and potentially, other component skills measured by ReadBasix) is moderated by the decoding threshold, intervention gains are more likely to be detected if students are above the decoding threshold.

ETS was contracted by Reading Reimagined to replicate decoding threshold analyses in tandem with its EFLR projects. The replication is necessary for two reasons. First, previously published decoding threshold results were based on students taking RISE assessments in a research project setting. RISE has since been commercialized by Capti,[2] an ETS approved vendor, as ReadBasix. Although most of the items themselves remain the same, there are differences between RISE and ReadBasix in the construction of the test forms, test length, and the scoring scale. Second, a growing number of students are using ReadBasix as a commercialized product, and the commercial sample may be different from the previous research sample in existing decoding threshold publications. Collectively, both the old and new

samples provide a larger data set to evaluate whether the prior findings generalize with a broader sample.

To address these issues, Reading Reimagined asked ETS to conduct two replication studies. Study 1 involves the replication of findings reported in Wang et al. (2019) based on a larger data set that has become available since 2019. Specifically, Study 1 includes (a) the identification of the decoding threshold and (b) an examination of the development trajectories of reading comprehension for students below and above the decoding threshold. Extending beyond prior research, Study 1 also explores the differences between students below versus above the decoding threshold, including their performance on other ReadBasix subtests—Vocabulary, Morphology, Sentence Processing, and Basic Reading Efficiency—and growth trajectories on these subskills.

Study 2 is designed to replicate Wang et al. (2020), which was designed to understand the behavioral differences between students scoring above and below the decoding threshold that help understand poor decoding skills.

Study 1a: Location of the Decoding Threshold

The goal of Study 1a is to identify the location of the decoding threshold on the new ReadBasix scoring scale. Compared to the original RISE scoring scale, which ranged between 190 and 310 with a mean of 250 and standard deviation of 15, the ReadBasix scale has the same effective range, the same mean, but a different standard deviation of 25. In order to use ReadBasix to identify students who face challenges with decoding, it is necessary to reestablish the cutoff point for the decoding threshold as reported in Wang et al. (2019, 2020).

Method

Sample

The sample included 167,403 students from Grades 3–12, *M* = Grade 7.2, Median = Grade 7, *SD* = 1.5 grade levels. The students came from three sources. Subsample 1 was from an urban school district on the U.S. east coast (*n* = 159,851). Subsample 2 was from other parts of United States and aimed to improve the national representativeness (*n* = 4,329). The two subsamples completed the RISE on the ETS research platform as part of the Reading for

Understanding IES grant (R305F100005). Subsample 3 consisted of students who completed ReadBasix on the Capti platform (n = 3,223) between 2020 and 2022. Sample weights were obtained such that each subsample and grade level have equal contribution to subsequent analysis (this was to downweigh the influence of Subsample 1 since this large sample was from a single school district). The sample and sampling weights are the same as those used to derive ReadBasix score norms, which are used to transform student scores into grade level specific percentile values.

Measures

All students in the sample completed both the WRD subtest and the RC subtest of RISE or ReadBasix. In WRD, students see one letter string at a time and need to decide whether it is (a) a real word, (b) not a word, or (c) an alternative spelling of a real word. In RC, students read passages that are about 200 words long and answer multiple choice questions that evaluate their understanding. The test length of WRD in the RISE test forms was about 50 items and in the ReadBasix it was about 30 items. The test length of RC in both RISE and ReadBasix was about 20 items, which were nested under four independent reading passages. Items on RISE and ReadBasix were drawn from the same item pool. RISE has nine test forms, and ReadBasix has 12 test forms. The same item across test forms was treated to have the same item parameters under the 2-parameter logistic (2PL) item response theory (IRT) model (Baker & Kim, 2004). Because of this, when students took different test forms, their ability estimate (θ) on the same subtest was on the same scale and thus comparable. The IRT marginal reliability of the WRD subtest ranged between .81 and .92, and the IRT marginal reliability of the RC subtest ranged between .70 and .85 across grade levels (Sabatini et al., 2019). Because the WRD subtest in the ReadBasix test forms was shorter than that in the RISE forms (30 vs. 50), we used the Spearman-Brown prophecy formula (de Vet et al., 2017) to estimate the reliability of the shortened WRD subtest in ReadBasix forms, and the estimated reliability was between .71 and .87.

The original RISE scoring scale and the ReadBasix scoring scale have the same mean M = 250, the same range [190, 310], but different standard deviation: Standard deviation on the RISE scoring scale was 15 whereas standard deviation on the ReadBasix scoring scale was 25. In

this replication study, all students' WRD and RC performance were scored on the ReadBasix scale.

Analysis

The sample was randomly divided into eight equal-sized batches (*n* = 20,925 or 20,926 across the batches). We conducted broken-line regression with the *R* package *lm.br* (Adams, n.d.) to examine the relation between decoding and reading comprehension performance on each of the eight batches. We used this approach because we were unable to run the analysis with a single batch due to limited computer memory.

Compared to linear regression, which estimates a single slope for the relation between two variables, broken-line regression estimates two slopes and the location of a threshold point where the slope changes. As such, whereas linear regression estimates two parameters—an intercept and a slope—broken-line regression estimates four parameters—an intercept, two slopes (Slope 1, Slope 2) and a threshold point. In the *lm.br* package, three types of broken-line regression can be specified: threshold-line (TL) relationship, which constrains Slope 1 to be zero; line-threshold (LT) relationship, which constrains Slope 2 to be zero; and line-line relationship, which does not constrain the slope. In the first two specifications, the number of parameters is three (instead of four), since one slope is constrained to be zero.

In this study, considering both theoretical and practical reasons, and to be consistent with Wang et al. (2019), we specified the TL relationship to identify the location of decoding threshold separately for each of the eight batches. We focused on two aspects of the broken-line regression results. First, we examined whether there was a significant TL relationship between WRD and RC. The *R* package *lm.br* provides significance level testing by comparing model fit of a two-slope model to a single slope model. Second, we identified the location on the WRD scale where the relationship between WRD and RC changed. The method employed by the *lm.br* package examined the distribution of the threshold point's conditional likelihood ratio. We then compared this location to Wang et al. (2019).

Examining Model Fit

The broken-line regression parameters were averaged across the eight batches to derive a function to predict RC with WRD. The prediction function was applied to all the student data, resulting in a predicted RC score for each student, $\widehat{RC}_{br}$ (br is short for broken-line regression). We compared students' actual RC scores with their $\widehat{RC}_{br}$ to examine residuals and then calculated AIC_{br} (Akaike, 1973) as an indicator of model fit. Separately, we also used simple linear regression to predict students' RC scores with WRD scores, $\widehat{RC}_{lr}$ (lr is short for linear regression) and used a similar procedure to calculate AIC_{lr}. We compared the two AIC values to show whether the broken-line regression model provided a better model fit than a single slope model.

Results

Figure 1 shows a scatterplot between students' WRD scores and RC scores.

Figure 1. Scatterplot of Decoding and Reading Comprehension Performance

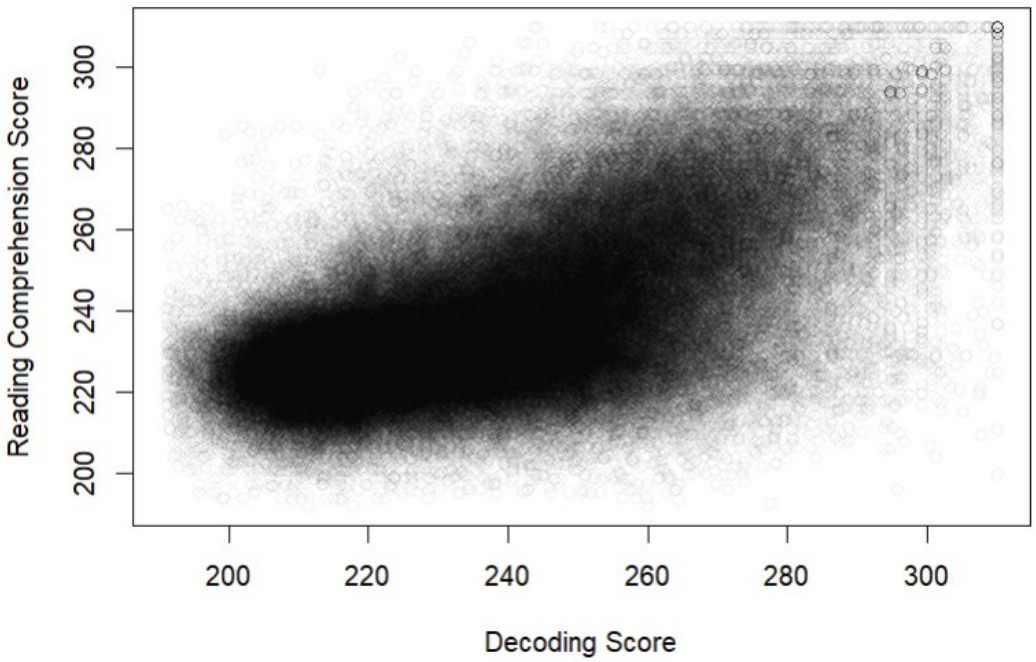


Table 1 summarizes the threshold relations identified from each of the eight batches. There was a significant threshold relation in all eight batches, $p < .01$. Across the eight batches, ***the average threshold point was at ReadBasix WRD score = 225.5***. On the ReadBasix scoring

scale, the location of the decoding threshold is at −1 *SD* from the mean (i.e., [250 − 225.5]/25 = −.98). In comparison, in Wang et al. (2019), the location of the threshold point was estimated to be at decoding = 235 on the RISE scale. Because the *SD* on the RISE scale was 15, the location of decoding threshold in that paper was also −1 *SD* from the mean (i.e., [250-235]/15 = -1). ***Thus, the location of the decoding threshold remains the same for the ReadBasix test when expressed in standardized scores.***

Table 1. Threshold Relations Based on ReadBasix Script and Norming Sample

Batch	Threshold*	Intercept	Slope1	Slope2
1	220.7	226.3	0	0.66
2	227.3	227.7	0	0.70
3	228.1	228.5	0	0.76
4	222.6	225.3	0	0.69
5	224.7	227.9	0	0.71
6	230.2	229.2	0	0.78
7	224.5	226.8	0	0.74
8	226.0	227.6	0	0.68
Mean	225.5	227.4	0	0.71

Note. Slope 1 was constrained to be zero to identify the location of threshold point.
*$p < .01$.

Model Fit

Using the averaged parameters of Table 1, the prediction function using broken-line regression was

$$\widehat{RC}_{br} = 227.4 + 0.71 * \max(WRD - 225.5, 0).$$

We also used simple linear regression to predict RC with WRD, and the regression equation was

$$\widehat{RC}_{lr} = 94.4 + .61 * WRD.$$

The AIC for the broken-line regression model was 1,392,324, whereas the AIC for the linear regression model was 1,398,943, which is 6,629 higher. Thus, the broken-line regression model had substantially better model fit than the simple regression model.

Using the ReadBasix norm, we estimated the percentages of students who were below the decoding threshold (<=225) at each grade level: Grade 3 – 39%, Grade 4 – 28%, Grade 5 – 25%, Grade 6 – 24%, Grade 7 – 21%, Grade 8 – 16%, Grade 9 – 11%, Grade 10 – 8%, Grade 11 – 6%, and Grade 12 – 5%. The percentages were lower than Wang et al. (2019) due to weighting: whereas the sample in Wang et al. (2019) was from a single urban school district on the East Coast of United States, the sample that was used to create ReadBasix performance norms were more diverse, including a national sample.

Study 1b: Impact of Decoding Threshold on Other Reading Skills and Development

The goal of Study 1b is to examine the impact of poor decoding skills on other reading subskills that are measured by ReadBasix (formerly known as RISE). In addition to WRD, ReadBasix has the following subtests: Vocabulary, Morphology, Sentence processing, Reading Efficiency, and RC. Study 1b is a longitudinal analysis of students' performance on these RISE subtests by students' initial decoding status: those above the decoding threshold versus those below.

Method

Sample

In each of the four fall semesters between 2011 and 2014, ETS administered the RISE battery to students in an urban school district on the East Coast of the United States. As a result, a total of 17,133 students provided multi-year longitudinal RISE performance: 11,705 students completed the test in 2 of the 4 years, 4,595 students in 3 of the 4 years, and 833 students every year during the 4 years. The distribution of grade levels when these students first took the RISE battery was 3,820 students from Grade 5; 6,448 from Grade 6; 3,646 from Grade 7; 1,947 from Grade 8; and 1,272 students from Grade 9. Efforts were taken in form assignment so that when students took the RISE test form again, they were usually administered a different test form.

Analysis

Similar to Study 1a, student raw responses on RISE items were scored on the ReadBasix scale. Each student's decoding status was determined by their lowest decoding score: If the decoding score was below 225, the student was determined to be below the decoding threshold; otherwise, the student was above the decoding threshold. In other words, students were categorized to be below the decoding threshold if they had scored below 225 on WRD in any test administration during the longitudinal data collection. The decision to categorize students this way was due to several considerations. First, this decision was consistent with Wang et al. (2019), the study to be replicated. Second, the decision was to simplify the longitudinal model to avoid model convergence problems. Third, this decision would produce a more conservative estimate of the difference in growth trajectories between students below versus above the decoding threshold, compared to treating decoding status as a time variant covariate or categorizing decoding status based on the average or highest decoding score across the years.

A set of six random effects longitudinal models were used, one for each ReadBasix subtest: WRD, Vocabulary, Morphology, Sentence Processing, Basic Reading Efficiency, and RC. The *R* package *lmr4* (Bates et al., 2015) was used for the analysis. The dependent variables of these models were the corresponding subtest score at a time point (year). The independent variables were time (with year as the unit), students' initial grade level (their grade level the first time they took the RISE battery), and their decoding status (below vs. above the threshold). The effect of time was allowed to vary between students as a random effect.

Because we were interested in understanding whether the growth trajectories differed between students who had been below the decoding threshold and those who had not during the longitudinal data collection, the interaction between time and student group was evaluated. The growth model for students' performance on the RC subtest is represented by the following equation:

$$RC_{ij} = \gamma_{00} + \gamma_{01} \times Grade_i + \gamma_{02} \times Decoding_i + \zeta_{0i} + (\gamma_{10} + \gamma_{11} \times Decoding_i + \zeta_{1i}) \times Time_{ij} + \varepsilon_{ij}.$$

The fixed effect of Time represents score improvement per year. The Time variable was centered so that at the first time a student took the RISE battery, Time = 0. For example, if a student first took RISE in 2011 and again in 2013, the Time variable for the 2011 data point would be 0, and the 2013 data point would be 2. As a result of the centering procedure for the Time variable, the intercept of the model represents the estimated value of the subtest score when students took the test for the first time in the study period.

The Grade variable represents the grade level of a student when he/she first took the RISE test during the study period. The Grade variable was centered so that the values for Grades 5–9 are 0, 1, 2, 3, 4, respectively. Thus, the intercept of the model represents a typical fifth grader's score. The fixed effect of Grade reflects the expected gain in RISE subtest scores when a student moved up a grade level.

The Decoding variable was binary. Decoding = 1 if a student was below the decoding threshold or 0 if above the decoding threshold. The coefficients of Decoding represent the fixed effects of being below the decoding threshold.

The statistical significance of each effect was examined by comparing the model with the effect and a baseline model that was created by removing the effect, using chi-square goodness of fit. A significant effect was indicated by a significant improved model fit over the baseline model based on the chi-square statistic.

Results

Table 2 provides a summary of fixed effects of the six models. All fixed effects are statistically significant at $p < .01$. To illustrate how to interpret results in this table, here we use the Vocabulary subtest as an example: The intercept is 233.5, meaning that Grade 5 students who were above the decoding threshold had an average Vocabulary score of 233.5 when taking the first RISE battery. The effect of Time (in years) was 5.5, indicating that on average, Grade 5 students' Vocabulary scores were expected to improve by 5.5 points after a year of instruction.

The effect of being below the decoding threshold was −17.3, which means that Grade 5 students who were below the decoding threshold had a Vocabulary score that was on average 17.3 points lower than Grade 5 students who were above the decoding threshold during the

first RISE test. The effect of Grade was 4.2; thus, Grade 6 students who were above the decoding threshold on average had an initial Vocabulary score 4.2 points higher than Grade 5 students, which was 233.5 + 4.2 = 237.7.

Finally, the interaction between Time and threshold group indicates how scoring below the decoding threshold resulted in slower growth. For Grade 5 students who were above the decoding threshold, the annual growth rate in Vocabulary was 5.5 points; this reduced to 5.5 – 3.2 = 2.3 points for Grade 5 students who were below the decoding threshold.

Table 2 summarizes the visualizations of effects in Figures 2−6 to demonstrate how being below the decoding threshold impacts other reading subskills and their development. Relevant to the aims of these analyses, ***these results demonstrate that (a) inadequate decoding skills are associated with low performance in other reading subskills cross-sectionally and (b) inadequate decoding skills predict slower growth in other reading subskills longitudinally.***

Table 2. Fixed Effects of Decoding Threshold Status on Intercept and Growth of Subskills

	Subskills (Dependent variables)				
Variable	Vocabulary	Morphology	Sentence	Efficiency	Comprehension
Intercept	233.5	235.5	238.4	234.4	238.8
Poor Decoding	-17.3	-19.6	-17.0	-17.4	-14.8
Time (year)	5.5	5.0	3.2	5.0	3.4
Time x Poor Decoding	-3.2	-2.6	-1.6	-2.9	-2.2
Grade	4.2	4.3	3.8	3.7	3.2

Note. Poor Decoding represents the effect of those below the decoding threshold relative to those above the decoding threshold. All fixed effects are statistically significant at $p < .01$.

Figure 2. Diverging Growth Trajectories in Vocabulary Scores of Students by Decoding Threshold Status

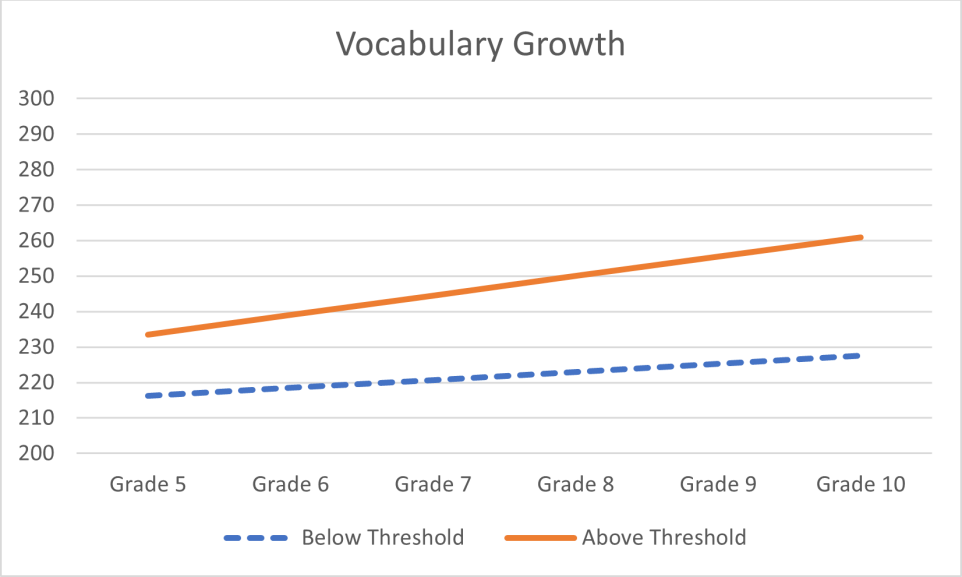


Figure 3. Diverging Growth Trajectories in Morphology Scores of Students by Decoding Threshold Status

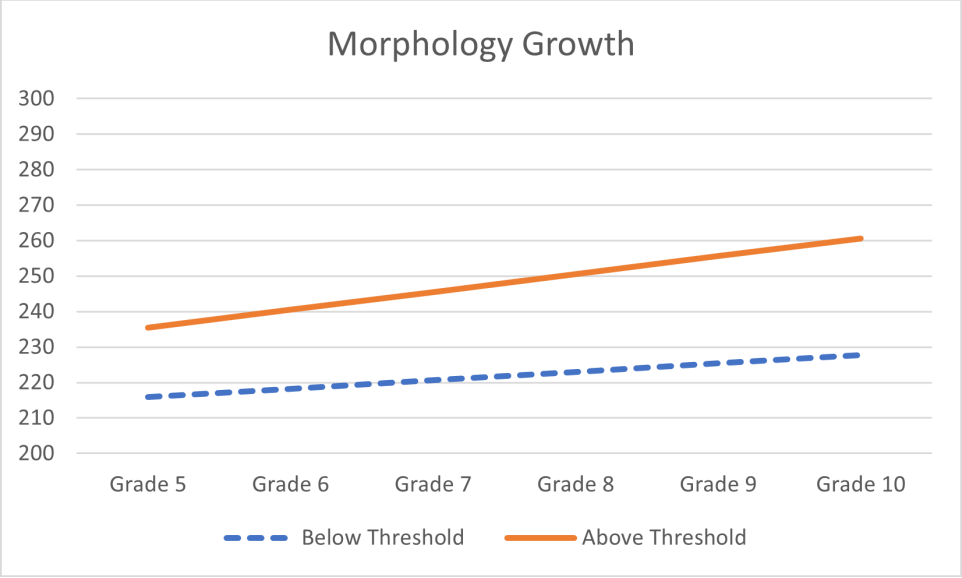


Figure 4. Diverging Growth Trajectories in Sentence Processing Scores of Students by Decoding Threshold Status

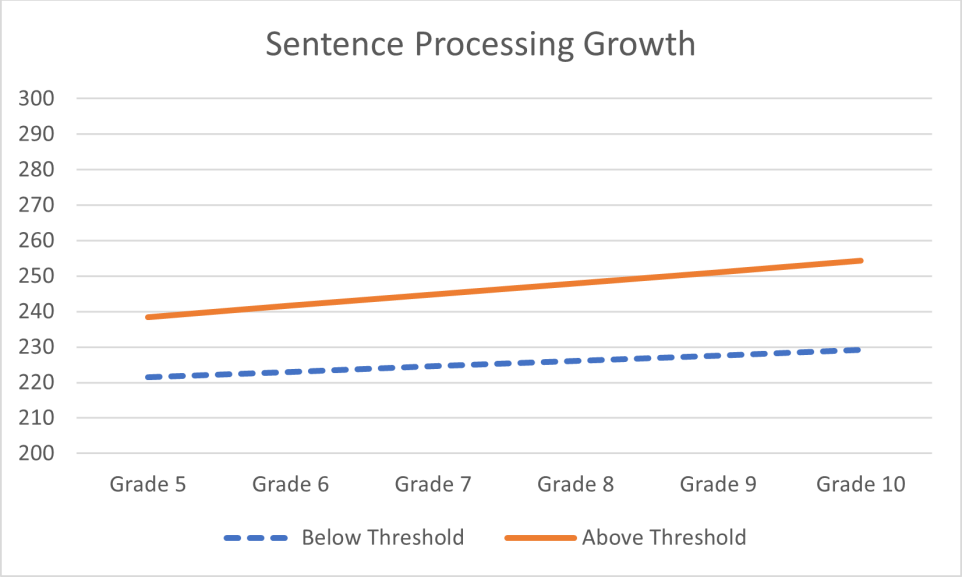


Figure 5. Diverging Growth Trajectories in Reading Efficiency Scores of Students by Decoding Threshold Status

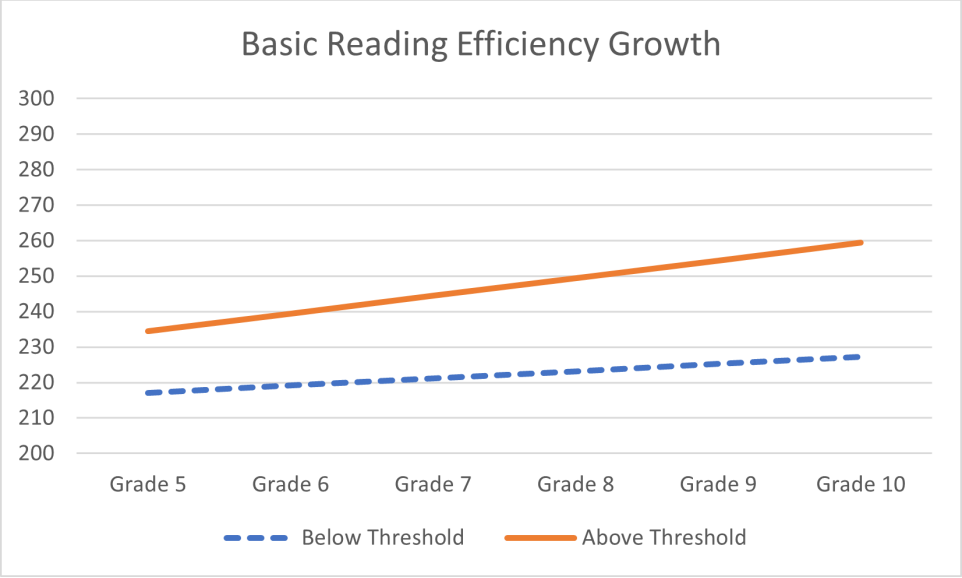
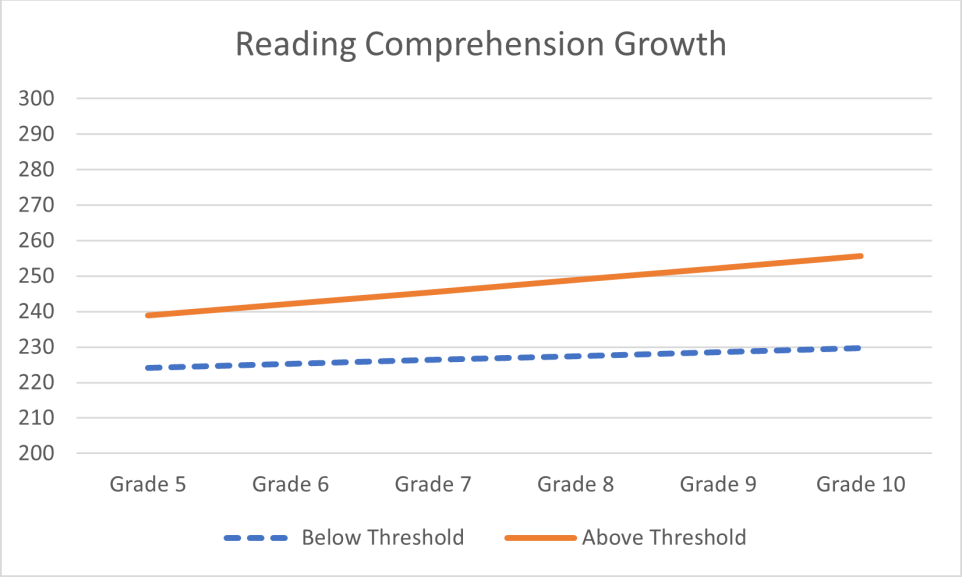


Figure 6. Diverging Growth Trajectories in Reading Comprehension Scores of Students by Decoding Threshold Status



Comparison With Existing Results

In Wang et al. (2019), students' growth trajectory on the RC subtest of ReadBasix (RISE) was explored by decoding threshold status. A comparison of results is provided in Table 3. Specifically, in Wang et al. (2019), the intercept was 247 on the RISE scale, which was equivalent to 245 on the ReadBasix scale. This intercept is higher than what we found in Study 1b (239). This finding indicates that students in Study 1b who were above the decoding threshold had on average lower RC scores than students in Wang et al. (2019). This divergence in results indicates sample differences.

In contrast, in both studies, students who were below the decoding threshold had almost identical RC scores: 225 in Wang et al. (2019) and 224 in Study 1b (225 is obtained by 245 – 20, and 224 = 239 – 15). This convergence supports the robustness of the location of the decoding threshold (i.e., 225 on the ReadBasix scale).

The annual growth rate for students above the decoding threshold, as reported in Wang et al. (2019), was 2.8 on the RISE scale or 4.7 on the ReadBasix scale. This rate is higher than what was found in Study 1b (2.4). Thus, for students above the decoding threshold, the Study 1b sample had both lower average RC scores and slower growth rate in RC. In contrast, for

students below the decoding threshold, their growth rate in RC was again well aligned in both studies, 1 point per year in Wang et al. (2019; i.e., 4.8–3.8) and 1.2 points in Study 1b (3.4–2.2).

To summarize the comparison of growth modeling in the current study and Wang et al. (2019), ***students in the current sample were on average of lower performance than those in Wang et al. (2019) and showed a slower growth rate. However, for students below the decoding threshold, the two studies produced comparable results.***

Table 3. Comparing Intercepts and Growth Rates in Reading Comprehension Across Studies

Variable	Wang et al. (2019) on RISE scale	Wang et al. (2019) on ReadBasix scale	Study 1b on ReadBasix scale
Intercept	247	245	239
Below Decoding Threshold	-12	-20	-15
Time (year)	2.9	4.8	3.4
Time x Below Decoding Threshold	-2.3	-3.8	-2.2
Grade	2.8	4.7	3.2

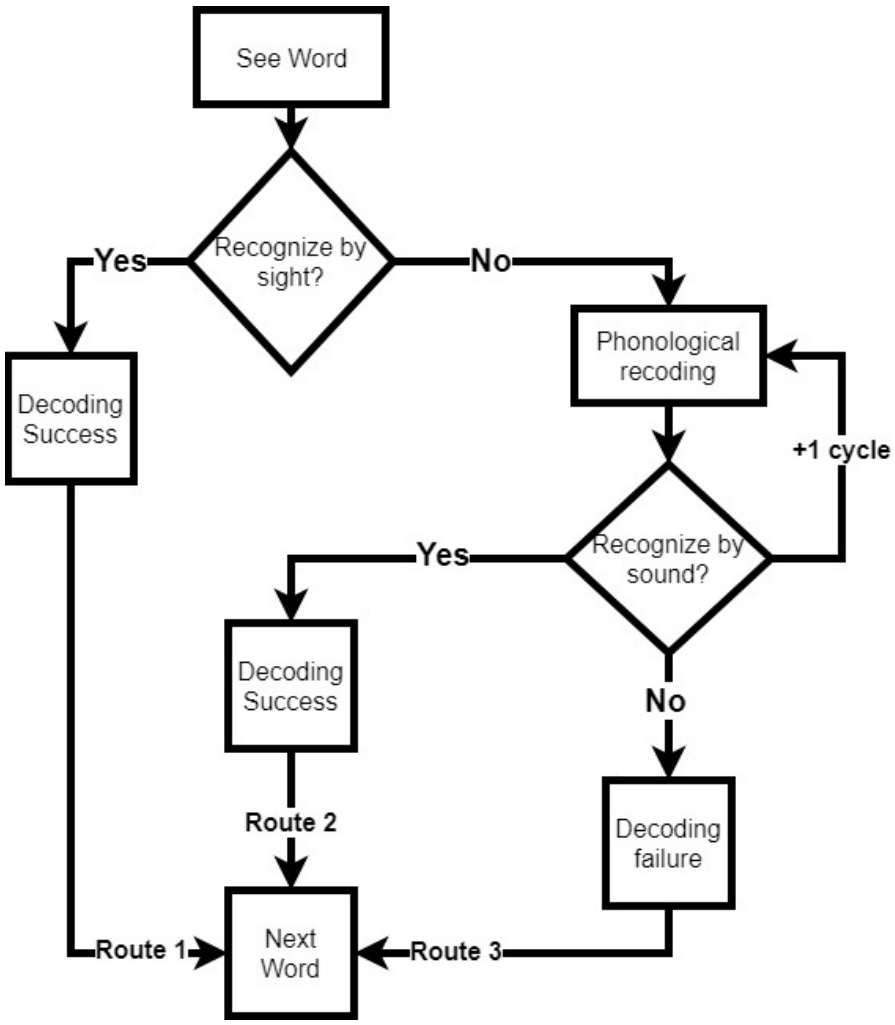
Study 2. Understanding Poor Decoding Skills From Decoding Behavior in ReadBasix

Whereas Study 1 identifies the decoding threshold (Study 1a) and illustrates the impact of being below the decoding threshold on other reading subskills (Study 1b), Study 2 explores the mechanism of poor decoding skill by examining test-taking behavior on ReadBasix WRD. Study 1b has shown that students who were below the decoding threshold also had slower growth rate in decoding in subsequent years. To understand why some students have developed poor decoding skills, Study 2 compares the time spent on different types of decoding items by student group (i.e., above vs. below the decoding threshold), which is a replication of Wang et al. (2020).

Wang et al. (2020) proposed the "developmental decoding process model" to account for poor decoding development (Figure 7). The model specifies three possible decision-making routes when a reader decodes a word. Route 1 shows the reader recognizes a word at first

sight. Route 2 shows the reader recognizes a word after engaging in some phonological recoding, that is, after sounding out the word with some effort before recognizing it. In contrast, Route 3 shows failed decoding, meaning that the reader fails to recognize a word. The developmental decoding process model posits that the time students spend trying to decode unfamiliar words as they encounter them, a process represented by the phonological recoding cycle in Figure 7, contributes to decoding skill development. This model predicts that (a) students who are below the decoding threshold would spend shorter time than peers on this phonological recoding process, and (b) the time students spend on phonological recoding predicts the growth rate in decoding.

Figure 7. Decision Making Process During a Decoding Task



The three routes can be represented by the three item types of the ReadBasix WRD subtest. In the test, students see one letter string at a time, and they need to decide if it is a real word, a pseudo-homophone, or a nonword. Making a correct selection on each of the three item types represents Routes 1, 2, and 3, respectively. Specially, each nonword item forces the test taker to go through Route 3 before they can reach the decision that it is a nonword (i.e., decoding failure). Logically, decoding success takes fewer steps and less time than decoding failure—if one phonological recoding cycle leads to word recognition, the test taker simply needs to make a selection; if, however, the test taker does not recognize the word after one phonological recoding cycle, they should still try some other variants of pronunciation by engaging in more phonological recoding cycles before they can be confident that the item is a nonword. Conversely, if test takers actively engage in the phonological recoding cycles, they should have longest response times when making a correct selection on nonword items and shortest response times when making a correct selection on real word items, with pseudo homophones in between. These hypotheses were tested in Study 2.

Method

Sample

The sample of Study 2 consisted of 14,498 ReadBasix test takers on the Capti platform in Grades 3–12 from June 2020 to January 2023 who took ReadBasix WRD. The great majority (70%) of these students took ReadBasix during the Fall semester of 2022. The mean grade level of this sample was 6.8, median was at Grade 7, and the standard deviation of grade levels was 2.1. Most of this sample (53.5%) was from Grade 6 (n = 3,770) and Grade 7 (n = 3,991).

Procedure

Students' WRD scores and time spent on each item were extracted from Capti's ReadBasix platform.

Analysis

Students' decoding status was determined by their decoding score: If they scored below 225, they were *below the decoding threshold* and, otherwise, *above the threshold*. Following the same procedure used in Wang et al. (2020), we only analyzed students' response times

based on correct responses. This is because the time spent on incorrect responses no longer reflects the processing time on the corresponding decision route (Figure 7). We used mixed design ANOVA to explore the effects of decoding status (between subjects) and word type (within subjects) on item time. The ANOVA analysis was conducted on SPSS 29.

Results

Out of the 14,498 students in this sample, 1,243 students did not provide any correct response on at least one of the three word types (the three item types were about evenly distributed across the 30 WRD items on each form), leading to 8.6% missing data (the missing rate in Wang et al., 2020 was 5.6%). The following results focus on the 13,255 students who have provided correct responses on all three item types.

Students' mean ReadBasix WRD scores were 245, *SD* = 17.5. The number of students below the decoding threshold was 1,945, or 14.7% of the sample (compared to 15.2% in Wang et al., 2020). Mixed design ANOVA revealed a significant interaction between word type and decoding status, $F(2, 26506) = 317.4$, $p < .01$, $\eta_p^2 = .023$. As shown in Figure 8, the significant interaction reflects the fact that whereas students above the decoding threshold spent longer time on nonword than pseudo-homophone items, students below the decoding threshold spent a similar amount of time on pseudo-homophones and nonwords.

Comparison With Existing Results

The current study replicated Wang et al. (2020) in two ways: (a) the significant interaction between item type and student group and (b) the significant main effect of item type. As shown in Figure 8 and Table 4, in both studies, ***students above the decoding threshold (labeled as normal decoders in the table) spent the longest time on nonword items, whereas students below the decoding threshold (labeled as poor decoders in the table) did not***. These results are consistent with the hypothesis that poor decoders do not engage in extra phonological recoding practice (Figure 6) when they encounter a word they do not recognize.

Figure 8. Item Time (in Milliseconds) by Student Group and Item Type

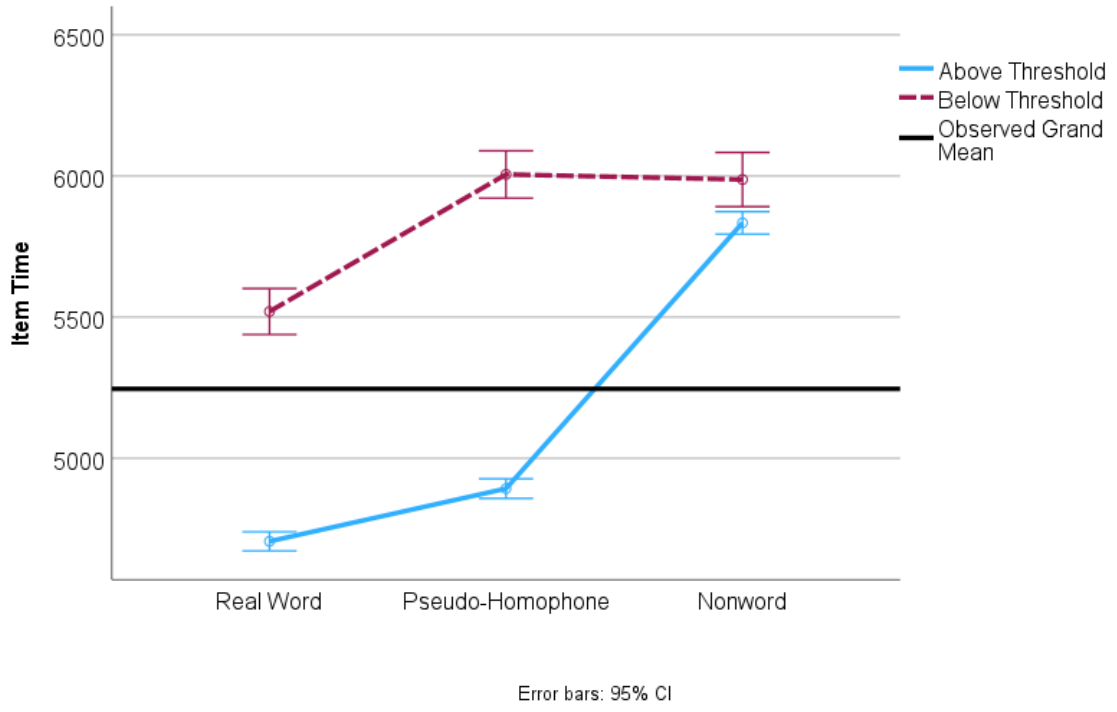


Table 4. Response Times on Three Types of Decoding Times by Study (in Seconds)

	Wang et al. (2020) *N* = 902 students		ReadBasix 2023 *N* = 13,255 students	
Item type	Normal decoders	Poor decoders	Normal decoders	Poor decoders
Real word	2.4	2.8	4.7	5.5
Nonword	3.8	3.0	5.8	6.0
Pseudo-homophone	2.8	3.2	4.9	6.0

Differences in Item Time Across Studies

The current sample on average spent longer time on WRD items than the sample in Wang et al. (2020) across the board (i.e., item type and student group). In the replication study sample, students above the decoding threshold spent between 2.0s and 2.3s longer, and students below the decoding threshold spent between 2.7s and 3.0s longer than the corresponding student groups in the 2020 study across the three item types (Table 4).

One factor that might have contributed to this difference in average time per item is the difference in test length. Students in Wang et al. (2020) took the RISE test forms, where the number of WRD items was about 50. In contrast, when students take the ReadBasix test forms, the number of WRD items was about 30. It is possible that students spent less time on each item in a longer test. Another factor might have to do with difference in the student samples. In Wang et al. (2020), students' average WRD score was 249 (*SD* = 13) on the RISE scale, which is about 248 (*SD* = 22) on the ReadBasix scale; in comparison, students in the current sample had an average Decoding score of 245 (*SD* = 18.5), which is slightly lower than the previous sample. The ability differences might have also contributed to differences in response times. Additionally, differences in how response time was captured between the RISE research platform and the ReadBasix commercial platform might also have played a role. For example, if one platform starts counting the response time of an item as soon as the test taker has finished the previous item, thus including the loading time of the item in response time, whereas another platform only starts counting the response time once the item is fully loaded, the former will result in longer response time.

Implications for Reading Reimagined and Future Directions

The new AERDF-funded studies replicated the following key findings of previously published results. First, the location of the decoding threshold is replicated in Study 1a on the ReadBasix scoring scale, which is at Decoding score = 225. Reading Reimagined's research projects and other programs should use this value to help interpret students' decoding performance.

Second, the impact of scoring below the decoding threshold on reading comprehension is replicated in Study 1b. Scoring below the decoding threshold is associated with stagnant growth in ReadBasix RC. Additionally, this relationship also generally applies to other ReadBasix measures including vocabulary, morphology, sentence processing, and basic reading efficiency. These findings suggest that inadequate decoding skills might become a bottleneck for development in other reading skills. This possibility is consistent with the self-teaching hypothesis (Share, 1995), which posits that successful decoding provides the developing reader with opportunities to learn the spelling-meaning connection of new words. Developing readers

who are below the decoding threshold have fewer opportunities to engage in this self-teaching process. In time, they will be left farther behind compared to peers (Figures 2–6). The implication is that identifying students who are below the decoding threshold early and providing them with effective intervention to raise them above the decoding threshold will likely facilitate their growth in other subskills of reading. This implication is consistent with the EFLR program's emphasis on foundational reading skills.

Third, Study 2 replicated the different behavioral patterns between students above versus below the decoding threshold across different types of decoding items. As a group, students scoring below the decoding threshold likely did not spend the extra time trying to perform phonological recording (Figure 6). This behavior is different from students who were above the decoding threshold. This finding speaks to the value of analyzing item level response time data in understanding students' decoding behavior, which has been shown to predict decoding growth (Wang et al., 2020). Although item level response time data are not yet reported to standard ReadBasix users, research users are welcome to contact the ETS team for assistance with such analysis.

Limitations and Future Directions

The original decoding threshold findings as reported in Wang et al. (2019, 2020) were based on a longitudinal data collection from 2011 through 2016, a span of five years. The data collection was part of a research study designed to collect longitudinal data from schools who were paid for their participation. These participating schools followed standard procedures provided by the ETS research team when administering the RISE reading assessments to their students. In contrast, ReadBasix became operational around 2021, and ReadBasix users are customers who pay Capti to use the assessment based on their needs. For example, whereas RISE test takers usually finished all six subtests during each wave of data collection, ReadBasix test takers often took a subset of all ReadBasix subtests. Additionally, we do not have longitudinal data from ReadBasix test takers due to the limited number of years since ReadBasix became available and because teachers use ReadBasix in a more flexible way than in a longitudinal research study. Furthermore, students do not necessarily keep the same

ReadBasix login ID when they move up a grade level, which makes tracking students across years a challenge.

The lack of ReadBasix longitudinal data prevents us from replicating all previous decoding threshold findings. This includes (a) replication of the impact of being below the decoding threshold on RC growth with ReadBasix data (Study 1b) and (b) replication of the impact of time spent on nonword items on WRD growth (Study 2). Replication of these results require a larger scale empirical study with ReadBasix users, which we leave to the future.

The discovery of the decoding threshold also changes how ReadBasix users may use the assessment. For students who score below the ReadBasix decoding threshold, teachers, knowing that they may struggle in the RC subtest, may decide to not administer the RC subtest yet, so as to save instructional time and avoid unnecessary frustration. Similarly, for students who are known to have adequate decoding skills and are working on reading comprehension, their teachers may decide to skip the WRD subtest. Such testing decisions made by teachers may impact the observed relation between ReadBasix WRD and RC. We leave this to future research.

References

Adams, M. (n.d.). lm. br: *An R package for broken line regression.* The Comprehensive R Archive. http://cran.pau.edu.tr/web/packages/lm.br/vignettes/lm.br.pdf

Akaike, H. (1973). Maximum likelihood identification of Gaussian autoregressive moving average models. *Biometrika*, *60*(2), 255–265. https://doi.org/10.1093/biomet/60.2.255

Baker, F. B., & Kim, S.-H. (2004). *Item response theory: Parameter estimation techniques* (2nd edition). CRC Press.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

de Vet, H. C. W., Mokkink, L. B., Mosmuller, D. G., & Terwee, C. B. (2017). Spearman–Brown prophecy formula and Cronbach's alpha: Different faces of reliability and opportunities for new applications. *Journal of Clinical Epidemiology*, *85*, 45–49. https://doi.org/10.1016/j.jclinepi.2017.01.013

Gough, P. B., & Tunmer, W. E. (1986). Decoding, reading, and reading disability. *Remedial and Special Education*, *7*(1), 6–10. https://doi.org/10.1177/074193258600700104

Hoover, W. A., & Gough, P. B. (1990). The simple view of reading. *Reading and Writing*, *2*(2), 127–160. https://doi.org/10.1007/BF00401799

Mol, S. E., & Bus, A. G. (2011). To read or not to read: A meta-analysis of print exposure from infancy to early adulthood. *Psychological Bulletin*, *137*(2), 267–296. https://doi.org/10.1037/a0021890

National Governors Association Center for Best Practices & Council of Chief State School Officers. (2010). *Common Core State Standards Initiative: About the standards.* https://www.thecorestandards.org/about-the-standards/

O'Reilly, T., Sands, A., Wang, Z., Dreier, K., & Sabatini, J. (2019). *Curbing America's reading crisis: A call to action for our children* [Policy report]. ETS. https://www.ets.org/s/research/pdf/curbing-america-reading-crisis.pdf

Perfetti, C. A., & Hart, L. (2002). The lexical quality hypothesis. In L. Verhoeven, C. Elbro, & P. Reitsma (Eds.), *Studies in written language and literacy: Precursors of functional literacy*

(Vol. 11, pp. 189–213). John Benjamins Publishing. https://doi.org/10.1075/swll.11.14per

Sabatini, J., Weeks, J., O'Reilly, T., Bruce, K., Steinberg, J., & Chao, S.-F. (2019). *SARA Reading Components Tests, RISE forms: Technical adequacy and test design, 3rd edition* (Research Report No. RR-19-36). ETS. https://doi.org/10.1002/ets2.12269

Share, D. L. (1995). Phonological recoding and self-teaching: *Sine qua non* of reading acquisition. *Cognition*, *55*(2), 151–218. https://doi.org/10.1016/0010-0277(94)00645-2

Wang, Z., Sabatini, J., & O'Reilly, T. (2020). When slower is faster: Time spent decoding novel words predicts better decoding and faster growth. *Scientific Studies of Reading*, *24*(5), 397–410. https://doi.org/10.1080/10888438.2019.1696347

Wang, Z., Sabatini, J., O'Reilly, T., & Weeks, J. (2019). Decoding and reading comprehension: A test of the decoding threshold hypothesis. *Journal of Educational Psychology*, *111*(3), 387–401. https://doi.org/10.1037/edu0000302

Notes

[1] For more details, please see https://aerdf.org/programs/reading-reimagined/eflr/

[2] https://www.captivoice.com/capti-site/public/entry/readbasix