Hammill Institute on Disabilities

Assessment for Effective Intervention
2021, Vol. 47(1) 53–58

DOI: 10.1177/1534508420937801
aei.sagepub.com
SAGE

Addressing the Large Standard Error of Traditional CBM-R: Estimating the Conditional Standard Error of a Model-Based Estimate of CBM-R

Joseph F. T. Nese, PhD[1], and Akihito Kamata, PhD[2]

Abstract

Curriculum-based measurement of oral reading fluency (CBM-R) is widely used across the country as a quick measure of reading proficiency that also serves as a good predictor of comprehension and overall reading achievement, but it has several practical and technical inadequacies, including a large standard error of measurement (*SEM*). Reducing the *SEM* of CBM-R scores has positive implications for educators using these measures to screen or monitor student growth. The purpose of this study was to compare the *SEM* of traditional CBM-R words correct per minute (WCPM) fluency scores and the conditional *SEM* (CSEM) of model-based WCPM estimates, particularly for students with or at risk of poor reading outcomes. We found (a) the average CSEM for the model-based WCPM estimates was substantially smaller than the reported SEMs of traditional CBM-R systems, especially for scores at/below the 25th percentile, and (b) a large proportion (84%) of sample scores, and an even larger proportion of scores at/below the 25th percentile (about 99%) had a smaller CSEM than the reported SEMs of traditional CBM-R systems.

Keywords

reading/literacy, progress monitoring, statistical methods

Curriculum-based measurement of oral reading fluency (CBM-R) is widely used across the country as a quick measure of reading proficiency that also serves as a good predictor of comprehension and overall reading achievement. But CBM-R has several practical and technical inadequacies, including large measurement error, or standard error of measurement (*SEM*; Ardoin & Christ, 2009; Christ et al., 2012; Christ & Silberglitt, 2007; Poncy et al., 2005). As teachers use CBM-R data to screen for students at risk of poor reading outcomes and monitor student progress to inform instructional decisions (Christ & Coolong-Chaffin, 2007), the large *SEM* of traditional CBM-R can affect the interpretations and consequences of the assessment results, with implications for instructional decision-making. With a large *SEM*, observed scores do not necessarily reflect true scores; therefore, educators could make an inappropriate educational decision based on an observed score that is higher than the true score (e.g., decide a poor intervention is working and should continue) or an observed score that is lower than the true score (e.g., decide a good intervention is not working). For CBM-R measures to have meaningful consequential validity for educators, scores need to be sensitive to instructional change, and the smaller the *SEM* the better.

Standard Error of Measurement (SEM)

The *SEM* is a measure of precision of an assessment score, where the standard deviation of the measure is multiplied by the square root of one minus the reliability of the measure:

$$\text{SEM} = SD \times \sqrt{1 - \rho_{xx'}}.$$

The smaller the *SEM*, the more precise the score. The *SEM* is generally more useful than a reliability coefficient for assessment consumers (like teachers) because it informs the interpretation of scores. For example, the *SEM* can be used to generate confidence intervals (CI) around reported scores; that is, the range, given a specific degree of certainty (usually 95%), within which a student's "true score" is contained. A range of about ± 2 *SEM* around a reported score provides a 95% CI that contains a student's true score.

[1]University of Oregon, Eugene, USA
[2]Southern Methodist University, Dallas, TX, USA

Corresponding Author:
Joseph F. T. Nese, University of Oregon, 5262 University of Oregon, Eugene, OR 97403, USA.
Email: jnese@uoregon.edu

SEM of Traditional CBM-R Scores

Traditional CBM-R WCPM scores are reported with a single *SEM* using a classical test theory approach. That is, a single *SEM* statistic is reported at the grade level, such that all grade-level CBM-R WCPM scores are associated with the same *SEM*. The values of the *SEM* of traditional CBM-R measures have been reported to range from 5 to 20 WCPM (e.g., Christ & Silberglitt, 2007; Poncy et al., 2005), and although data with a *SEM* of 5 WCPM have been anecdotally described as "very good" (Christ et al., 2012, p. 365), a more realistic range is 6 to 12 WCPM (aimswebPlus, 2018; Alonzo & Tindal, 2009; Christ & Colleagues, 2015; University of Oregon, 2019). For example, the reported WCPM *SEM* ranges across early elementary grades for several CBM-R assessments are as follows: aimsweb, 6.28 to 9.58 WCPM (aimswebPlus, 2018); DIBELS 8, 7.12 to 11.23 WCPM (University of Oregon, 2019); easyCBM, 7.71 to 12.11 WCPM (Alonzo & Tindal, 2009); and FastBridge, 8.54 to 10.41 WCPM (Christ & Colleagues, 2015). It is important to note that the CBM-R *SEM* generally increases across grade levels for all CBM-R systems.

When measuring student progress, smaller SEMs become quite important. For example, the updated Hasbrouck and Tindal (2017) CBM-R WCPM norms reported an average fall CBM-R score of 59 WCPM, and an average spring score of 91 WCPM, for Grade 3 students at the 25th percentile (a reasonable percentile cut that a school may use in practice to define students who are at risk for poor reading outcomes and thus target for instructional supports). The average expected within-year growth for a Grade 3 student at the 25th percentile is 32 WCPM (91–59). Given a realistic *SEM* of 8 WCPM, we can estimate a 95% CI: $\pm 2 \times 8$ WCPM (*SEM*) $= \pm 16$, or a 95% CI of 32 WCPM around any score between 59 and 91. Thus, the 95% CI of 32 WCPM is equal to the expected growth for the entire year for an average at-risk student at the 25th percentile. The large CI is problematic when CBM-R measures are used to monitor student progress and to help make instructional decisions because a teacher cannot separate student learning from measurement error.

Conditional SEM (CSEM)

Estimates of the standard error (*SE*) at different score levels are referred to as conditional standard errors of measurement, or CSEM. Conceptually, the *SEM* is generally equivalent to the average CSEM for a given sample. According to the Standards for Educational and Psychological Testing (American Educational Research Association et al., 2014), the CSEM "can be much more informative than a single average *SE* for a population" (p. 39). This is true for CBM-R (and CBM in general) where educators use scores to screen for students at risk of poor reading outcomes, and modify instruction based on student progress data. Teachers and school teams analyze and evaluate student assessment data to inform educational decisions such as: *Based on universal screening data, is the student at risk of poor reading outcomes? Based on progress monitoring data, is the intervention working? If* "yes," *should instruction cease or continue? If* "no," *should instruction continue or be modified?*

In multitiered systems of support, students at risk of poor learning outcomes and targeted as candidates for intervention are often identified using a percentile cut score on a norm-referenced test (e.g., at/below 25th percentile). Thus, for progress monitoring data, decisions are concentrated in one area of the score distribution (e.g., at/below the 25th percentile), making the use of CSEM particularly valuable for CBM-R decision-making. The Standards for Educational and Psychological Testing (American Educational Research Association et al., 2014) also indicate that "if decisions are based on test scores and these decisions are concentrated in one area or a few areas of the scale score, then the conditional errors in those areas are of special interest" (p. 39). As instructional decisions are often based on CBM-R scores, and these decisions are generally concentrated in the bottom portion of the score distribution, the CSEM for students scoring at/below the 25th percentile are likely of special interest to teachers and researchers (Hasbrouck & Tindal, 2017; Nese et al., 2011). Thus, in this article we also examined the CSEM for students at/below the 25th percentile.

Model-Based Estimate of WCPM

The present study uses a model-based estimate of WCPM, based on a recently proposed latent-variable psychometric model of speed and accuracy for CBM-R data (Kara et al., 2020). The model-based CBM-R WCPM estimates are based on a two-part model that includes components for reading accuracy and reading speed. The accuracy component is a binomial-count factor model in which accuracy is measured by the number of correctly read words in the passage. The speed component is a log-normal factor model in which speed is measured by passage reading time. Parameters in the accuracy and speed models are jointly modeled and estimated. For a detailed description, please see Kara et al. (2020).

There are several advantages of the model-based WCPM estimates compared with traditional CBM-R WCPM scores. For example, CBM-R passages can be equated between multiple grade levels, placing the estimated WCPM scores on a common scale and making it especially useful for progress monitoring. In addition, the model-based WCPM estimates are on the same metric as traditional CBM-R scores (i.e., WCPM), which makes the scale scores immediately usable for teachers and reading specialists who are familiar with the WCPM expectations for their students. This study focuses on the fact that the *SE*s of the model-based WCPM scores can be computed for each observation with a single test administration.

Purpose. The purpose of this study was to compare the *SEM* of traditional CBM-R WCPM scores and the CSEM of the model-based WCPM estimates proposed by Kara et al. (2020), particularly for students with or at risk of poor reading outcomes. Our research questions are as follows.

Research Question 1: Is the average CSEM for the model-based WCPM estimates smaller than the reported SEMs of traditional CBM-R systems?
Research Question 2: What is the proportion of students that have a smaller CSEM than the reported SEMs of traditional CBM-R systems?
Research Question 3: Is the average CSEM of students at/below the 25th percentile for the model-based WCPM estimates smaller than the reported SEMs of traditional CBM-R systems?
Research Question 4: What is the proportion of students at/below the 25th percentile that have a smaller CSEM than the reported SEMs of traditional CBM-R systems?

Table 1. SEM of Traditional CBM-R Systems and Mean CSEM of Model-based WCPM Estimates.

	SEM		
CBM-R	Grade 2	Grade 3	Grade 4
aimswebPlus	7.78	7.46	8.40
DIBELS 8th Edition	7.84	9.59	9.63
easyCBM	–	9.73	–
FastBridge	8.54	8.54	10.41
	CSEM		
Model-based WCPM estimate	5.15	5.47	7.63

Note. The aimswebPlus represent the smallest SEM reported in the research report across fall, winter, and spring estimates. The model-based WCPM estimates were computed based on the Bayesian approach described in Kara et al. (2020). SEM = standard error of measurement; CBM-R = curriculum-based measurement of oral reading fluency; CSEM = conditional SEM.

Method

This study was part of a larger project to develop and validate a computerized assessment of CBM-R that uses (a) automated speech recognition to score students' readings and (b) an advanced psychometric model to overcome some of the inadequacies of traditional CBM-R, such as the large *SEM*. See https://jnese.github.io/core-blog/ for a full description of the larger project.

Data are from a study that compared the consequential validity properties of the computerized assessment and a traditional CBM-R for students in Grades 2 through 4 using a repeated measurement design across 2017–2018. See (https://jnese.github.io/core-blog/posts/2019-04-12-consequential-validity-study-procedures/) for a full description of the study, including more information on the measures and procedures.

Participants

The sample consisted of 4,084 CBM-R scores from 1,021 students in four schools and three school districts in the Pacific Northwest from 2017–2018; 343 were in Grade 2, 354 were in Grade 3, and 324 were in Grade 4. Approximately 50% were male; 22% were Latinx; 79% were White, 8% were multiracial, 7% were Hispanic, 4% were American Indian/Alaskan Native, and less than 1% were Asian, Black, or Native Hawaiian/Other Pacific Islander; 15% received special education services, and 8% received English Learner services.

Procedures

For each of the four measurement occasions (October 2017, November 2017, February 2018, May 2018), students read aloud online a randomly assigned, fixed set of 10 to 12 passages that were each 50 or 85 words in length (±5 words). Students were to read each passage in its entirety but were stopped after 90 s if they had not finished reading. An automatic speech recognition engine scored each reading, scoring each word as read correctly or incorrectly (accuracy) and recording the time duration to read the passage (speed). Model-based WCPM estimates (Kara et al., 2020) were based on these readings and data, such that a model-based WCPM score was estimated for each measurement occasion based on the 10 to 12 passages read.

Analyses

We estimated model-based WCPM and the associated CSEM for each student (based on the set of passages read) at each measurement occasion, and compared the CSEM estimates with the reported *SEM* for the following traditional CBM-R systems: aimswebPlus, DIBELS 8th Edition, easyCBM, and FastBridge. (Note that we examined the data by grade and measurement occasion and found no differences across occasions, so we report all results by grade.)

To address Research Question 1, we computed the mean CSEM by squaring the CSEM for each CBM-R score, computing the mean $CSEM^2$, and taking the square root of that mean.

$$\overline{\text{CSEM}} = \sqrt{\frac{\sum_{i=1}^{n} \text{CSEM}_i^2}{N}}.$$

To address Research Question 2, we report the proportion of CBM-R scores that have a smaller CSEM than the *SEM* of traditional CBM-R. For our purposes here, we used a conservative reference *SEM* = 8 WCPM for each grade, based on the reported SEMs of traditional CBM-R systems (see Table 1). We propose that if more than 50% of CSEM

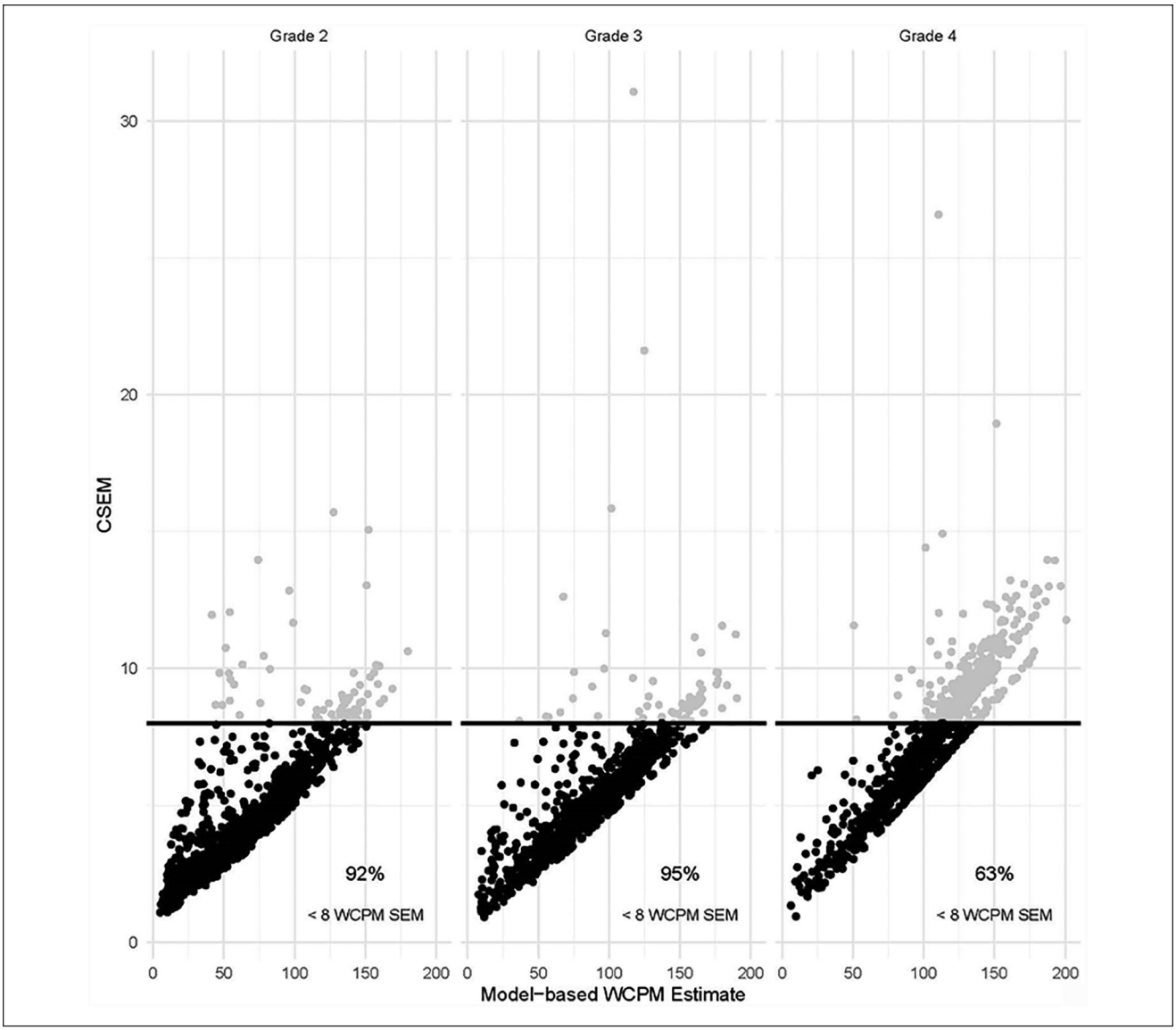

Figure 1. Model-based WCPM estimate.

estimates fall below 8 WCPM, this will provide reasonable evidence that the CSEM is of better quality than the *SEM*.

To address Research Questions 3 and 4, we repeated the analyses of the first two research questions using model-based estimated CBM-R scores at/below the 25th percentile by grade and measurement occasion.

All analyses were conducted in R (R Core Team, 2019), with the following packages: papaja (Aust & Barth, 2018), rio (Chan et al., 2018), and tidyverse (Wickham et al., 2019).

Results

Table 1 shows the *SEM* for Grades 2 through 4 reported by aimswebPlus, DIBELS 8th Edition, easyCBM, and FastBridge, as well as the mean CSEM for the model-based WCPM estimates. In response to Research Question 1, the average estimated CSEM for each grade was less than 8 WCPM; mean CSEM = 5.15 for Grade 2, 5.47 for Grade 3, and 7.63 for Grade 4. Thus, the average CSEMs of the model-based WCPM estimate were lower than the SEMs of the traditional CBM-R systems for all grades.

In response to Research Question 2, 92% of estimated CSEMs were less than of 8 WCPM for Grade 2, 95% for Grade 3, and 63% for Grade 4. These results were all well above the 50% criterion we established a priori. Across all grades, 84% of all 4,084 CBM-R scores had a CSEM less than 8 WCPM. Figure 1 shows the proportion of CSEM estimates that were below 8 WCPM, as well as the relation between the model-based estimated WCPM scores and the

Table 2. Mean CSEM of Model-Based WCPM Estimates and Percent Below 8 WCPM for Sample Students at/Below the 25th Percentile.

Grade	Unique Students	Number of CBM-R Scores	Mean CSEM	Percent < 8 WCPM
Grade 2	122	251	2.96	99.60
Grade 3	127	284	3.33	98.94
Grade 4	125	253	4.78	98.02

Note. Table results represent the total number of unique students and CBM-R scores at/below the 25th percentile across all four repeated measurement occasions. CSEM = conditional SEM; CBM-R = curriculum-based measurement of oral reading fluency.

CSEMs, which were positively correlated across grades (Grade 2, $r = .82$; Grade 3, $r = .78$; Grade 4, $r = .89$).

To address Research Questions 3 and 4, we selected CBM-R scores at/below the 25th percentile of model-based WCPM scores for each grade and measurement occasion. For these scores at/below the 25th percentile, Table 2 displays the estimated mean CSEM and the percentage of scores whose model-based WCPM score had a CSEM lower than 8 WCPM. In response to Research Question 3, the estimated mean CSEMs ranged from 2.96 to 4.78 WCPM, which was substantially smaller than the reported *SEM* of traditional CBM-R systems in Table 1. In response to Research Question 4, nearly all (98.02%–99.60%) scores at/below the 25th percentile had a CSEM less than 8 WCPM.

Discussion

Despite the prevalent use and practical application of CBM-R, the large *SEM* of traditional CBM-R leads to less accurate scores. The results of this study showed that the average CSEMs for the model-based WCPM estimates across grades were substantially smaller than the reported SEMs of traditional CBM-R systems (see Table 1), especially for students at/below the 25th percentile (see Table 2). We also found that a large proportion (84%) of CBM-R scores had a smaller CSEM than the reported SEMs of traditional CBM-R systems; this was especially true for students at/below the 25th percentile, of which about 98% of CBM-R scores had a CSEM smaller than the *SEM* of traditional CBM-R. We found positive correlations (r from .78 to .89) between the model-based estimated WCPM scores and the CSEMs (see Figure 1), which has implications for the applied use of the model-based WCPM scores, particularly for teachers using these measures to monitor the progress of students at the lower end of the score distribution who are at risk of poor reading outcomes. A lower CSEM implies an increased reliability of CBM-R scores, and thus provides educators with more accurate CBM-R scores that are more sensitive to instructional change than traditional CBM-R scores.

Limitations

It should be noted that the computerized administration of study passages was different than traditional administration of CBM-R passages, by design. Multiple passages were administered, intended to be read independently and in their entirety, that were calibrated, equated, and linked, and the model-based scoring was developed for automated speech recognition. Thus, the results presented here merit replication for purposes of reproducibility and generalization.

Conclusion

The results of this study provide preliminary but promising evidence that the model-based WCPM scale scores have a lower CSEM estimate compared with the *SEM* of traditional CBM-R WCPM scores, especially for students at risk of poor reading outcomes. Lower CSEM estimates make the model-based WCPM scores better suited for measuring CBM-R, both for screening and progress monitoring. A more precise screening score would help educators better identify students at risk of poor reading outcomes. Reducing the *SEM* of CBM-R progress monitoring assessments would help educators better determine student responsiveness to intervention. Thus, the model-based WCPM estimates could improve the consequential validity of CBM-R assessment systems by helping educators make better decisions to improve student reading outcomes.

Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305A140203 to the University of Oregon. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

References

aimswebPlus. (2018). *AimswebPlus: Efficacy research report*. Pearson.https://www.pearson.com/content/dam/one-dot-com/one-dot-com/global/Files/efficacy-and-research/reports/efficacy-assessment-reports/aimsweb-Plus-research-report.pdf

Alonzo, J., & Tindal, G. (2009). *Alternate form and test-retest reliability of easyCBM reading measures* (Technical report no. 0906). Behavioral Research and Teaching, University of Oregon. https://files.eric.ed.gov/fulltext/ED531558.pdf

American Educational Research Association, American Psychological Association, National Council on Measurement in Education, & Joint Committee on Standards for Educational

and Psychological Testing. (2014). *The standards for educational and psychological testing*. American Educational Research Association.

Ardoin, S. P., & Christ, T. J. (2009). Curriculum-based measurement of oral reading: Standard errors associated with progress monitoring outcomes from DIBELS, AIMSweb, and an experimental passage set. *School Psychology Review*, *38*, 266–283.

Aust, F., & Barth, M. (2018). *papaja: Create APA manuscripts with R Markdown*. https://github.com/crsh/papaja

Chan, C., Chan, G. C., Leeper, T. J., & Becker, J. (2018). *rio: A Swiss-army knife for data file I/O*. https://CRAN.R-project.org/package=rio

Christ, T. J., & Colleagues. (2015). *Formative assessment system for teachers: Benchmarks and norms for 2015–16*, Minneapolis, MN: FastBridge Learning, LLC. https://www.fastbridge.org/wp-content/uploads/2015/12/2015-16FastBridgeNormsandBenchmarksAllMeasuresFINAL.pdf

Christ, T. J., & Coolong-Chaffin, M. (2007). Interpretations of curriculum-based measurement outcomes: Standard error and confidence intervals. *School Psychology Forum*, *1*, 75–86.

Christ, T. J., & Silberglitt, B. (2007). Estimates of the standard error of measurement for curriculum-based measures of oral reading fluency. *School Psychology Review*, *36*(1), 130–146.

Christ, T. J., Zopluoglu, C., Long, J. D., & Monaghen, B. D. (2012). Curriculum-based measurement of oral reading: Quality of progress monitoring outcomes. *Exceptional Children*, *78*(3), 356–373.

Hasbrouck, J., & Tindal, G. (2017). *An update to compiled ORF norms* (Technical report no. 1702). Behavioral Research and Teaching, University of Oregon. https://files.eric.ed.gov/fulltext/ED594994.pdf

Kara, Y., Kamata, A., Potgieter, C., & Nese, J. F. (2020). Estimating model-based oral reading fluency: A Bayesian approach. *Educational and Psychological Measurement*. https://doi.org/10.1177/0013164419900208

Nese, J. F., Park, B. J., Alonzo, J., & Tindal, G. (2011). Applied curriculum-based measurement as a predictor of high-stakes assessment: Implications for researchers and teachers. *The Elementary School Journal*, *111*(4), 608–624.

Poncy, B. C., Skinner, C. H., & Axtell, P. K. (2005). An investigation of the reliability and standard error of measurement of words read correctly per minute using curriculum-based measurement. *Journal of Psychoeducational Assessment*, *23*(4), 326–338.

R Core Team. (2019). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

University of Oregon. (2019). *Dynamic indicators of basic early literacy skills 8th edition: Administration and scoring guide*. https://dibels.uoregon.edu

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., . . . Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, *4*(43), Article 1686. https://doi.org/10.21105/joss.01686