

Estimating Model-Based Oral Reading Fluency: A Bayesian Approach

Educational and Psychological
Measurement

2020, Vol. 80(5) 847–869

© The Author(s) 2020

Article reuse guidelines:

sagepub.com/journals-permissions

DOI: 10.1177/0013164419900208

journals.sagepub.com/home/epm



Yusuf Kara¹, Akihito Kamata¹ , Cornelis Potgieter^{2,3} and Joseph F. T. Nese⁴

Abstract

Oral reading fluency (ORF), used by teachers and school districts across the country to screen and progress monitor at-risk readers, has been documented as a good indicator of reading comprehension and overall reading competence. In traditional ORF administration, students are given one minute to read a grade-level passage, after which the assessor calculates the words correct per minute (WCPM) fluency score by subtracting the number of incorrectly read words from the total number of words read aloud. As part of a larger effort to develop an improved ORF assessment system, this study expands on and demonstrates the performance of a new model-based estimate of WCPM based on a recently developed latent-variable psychometric model of speed and accuracy for ORF data. The proposed method was applied to a data set collected from 58 fourth-grade students who read four passages (a total of 260 words). The proposed model-based WCPM scores were also evaluated through a simulation study with respect to sample size and number of passages read.

Keywords

oral reading fluency, speed–accuracy model, Bayesian estimation

¹Southern Methodist University, Dallas, TX, USA

²Texas Christian University, Fort Worth, TX, USA

³University of Johannesburg, Johannesburg, South Africa

⁴University of Oregon, Eugene, OR, USA

Corresponding Author:

Akihito Kamata, Center on Research and Evaluation, Southern Methodist University, Box 511, Dallas, TX 75275-0511, USA.

Email: akamata@smu.edu

Oral reading fluency (ORF) is defined as “the oral translation of text with speed and accuracy” (Fuchs et al., 2001, p. 239; Shinn et al., 1992), and there is strong theoretical support for ORF as an essential part of reading proficiency (LaBerge & Samuels, 1974; National Reading Panel, 2000; Perfetti, 1985). In addition, research has repeatedly established that ORF is an important indicator of reading comprehension and overall reading competence (e.g., Fuchs, 2004; Fuchs et al., 1988; Fuchs et al., 2001; Jenkins et al., 2003; Kim et al., 2010; Pinnell et al., 1995; Yovanoff et al., 2005). ORF can be assessed as a key component of reading ability and is a defining characteristic of good readers, while a lack of ORF is a common characteristic of poor readers (Hudson et al., 2005). Thus, the measurement of ORF is an important part of screening assessments for identifying students at risk of poor reading outcomes, as well as an important assessment of students’ response to reading intervention.

There are many standardized tests that used ORF assessments, such as the Reading Comprehension section of the Iowa Test of Basic Skills Battery (ITBS; Hoover et al., 2001), the reading test of the Colorado Student Assessment Program (CSAP; Colorado Department of Education, n.d.), the Stanford Achievement Test (10th ed.; SAT-10, Harcourt Brace, 2003), and the National Assessment of Educational Progress (NAEP). However, ORF has mostly been assessed as part of curriculum-based measurement (CBM), which is designed to measure students’ academic status and growth so the effectiveness of instruction may be evaluated. CBM is progressive measurement procedures that can be applied by teachers in order to monitor students achievements (in reading, writing, and math), and explore the possible needs for interventions and instructional modifications (Deno, 1985; Fuchs et al., 1988). Strong relations have been reported between ORF scores from CBM and scores from standardized reading tests (e.g., Crawford et al., 2001; Roehrig et al., 2008; Shinn et al., 1992; Valencia et al., 2010; Wood, 2006).

In a traditional CBM ORF administration, a student is given 1 minute to read as many words as possible in a grade-level text (approximately 250 words in length), while a trained assessor follows along and indicates on a scoring protocol each word the student reads incorrectly (Miura Wayman et al., 2007). If a student pauses for more than 3 seconds, the assessor prompts the student to continue and marks the word as read incorrectly. Student self-corrections are not marked as errors, but omissions are. After 1 minute, the assessor calculates words correct per minute (WCPM) by subtracting the number of incorrectly read words from the total number of words read. The WCPM score has been the most prevalent measure of ORF for decades and thought to be a good indicator of the overall reading competence with strong concurrent and predictive validity (Fuchs et al., 2001; Hasbrouck & Tindal, 2006).

An Improved ORF Assessment System

Despite prevalent use and practical application of ORF measures, the current standard ORF assessment has considerable practical and psychometric limitations which potentially make traditional ORF measures less reliable and valid. For example,

WCPM scores vary substantially across passages despite high correlations across passages of “grade-level” texts (Betts et al., 2009; Francis et al., 2008). Variability across ostensibly equivalent passages calls to question the appropriateness of using WCPM scores as indicators of student risk and as a mechanism to evaluate student growth. In addition, postequating is required to equate WCPM scores, which is likely sample specific. On the other hand, the inaccuracy of WCPM scores can be seen in the large standard errors (*SEs*; Christ & Silberglitt, 2007; Poncy et al., 2005), where the 95% confidence interval around a WCPM score is often larger than the magnitude of within-year expected growth (Hasbrouck & Tindal, 2017; Nese et al., 2013). This large error band is problematic when WCPM scores are used to monitor student progress and to help make educational decisions. Moreover, parallel-form or alternative-form reliability is required to estimate measurement errors, which again is sample-specific.

As a result of the psychometric limitations (and in addition to practical limitations not addressed here), an improved ORF assessment system has been developed to reduce these limitations (Nese et al., 2015). The improved ORF assessment system incorporates several modifications to the traditional ORF assessment. First, passage lengths are substantially shorter than in traditional ORF assessments, which are about 250 words: Medium length passages are approximately 50 words, and long passages are approximately 85 words. The intent for these shorter passages is for multiple passages to be administered to each student without increasing the burden to assessors or students in terms of time or demand. Second, unlike traditional ORF assessments, sufficient time is given, and students are intended to read each passage in its entirety. Third, the assessment delivery is computer-based, such that passages are presented to the student on a computer screen, and the system records students’ audio as they read the passage. As a result, the improved assessment system allows centralized scoring of the recorded reading audio by both human assessor and a speech recognition engine (Nese & Kamata, 2019). It also allows the assessment system to collect accuracy and time data at the word level, which can be aggregated at the sentence and passage levels.

Availability of word level data enables the estimation of ORF beyond the traditional WCPM scores. As part of the effort to establish an improved ORF assessment system, Potgieter et al. (2017) proposed a latent variable psychometric model to fit ORF data. The model proposed by Potgieter et al. parameterizes characteristics of passages and examinees with respect to speed and accuracy (described below); however, the model does not parameterize fluency or WCPM. Thus, the current study proposes a new model-based method to estimate WCPM scores based on the model parameters proposed by Potgieter et al. To our knowledge, no other work has demonstrated any model-based WCPM scores, the results of which contribute to the research literature on ORF assessment, and moreover, offer an improved ORF score to be applied in practice by educators to screen for and progress monitor students at risk of poor reading outcomes.

A Latent Variable Model for ORF

The latent variable model for ORF proposed by Potgieter et al. (2017) is a modification of a speed-accuracy model by van der Linden (2007). van der Linden's model is a two-part latent-variable model for speed and accuracy where each assessment item is a binary outcome (e.g., correct or incorrect). The speed component is a log-normal factor model, while the accuracy component is a three-parameter logistic item response theory (3PL IRT) model. The approach proposed by van der Linden uses a random-effects model, and parameter estimation is done using a hierarchical Bayes framework. The hierarchical Bayes approach is fairly common for this type of latent variable model; see Fox et al. (2007), Entink et al. (2009), and van der Linden et al. (2010). The two-part latent variable model proposed by van der Linden has grown in popularity in the literature, as the inclusion of response-time data can increase the information available for statistical modeling and inference (Ranger, 2013). In order to fit ORF assessment data, Potgieter et al. (2017) modified van der Linden's model by replacing the 3PL IRT model with a binomial count factor model for the accuracy part of the model, as accuracy of an ORF assessment is the count of words read correctly in a passage. The time duration to read a passage was modeled using the log-normal distribution, similar to the speed component in van der Linden's model. Below, the two components of the model proposed by Potgieter et al. are described in detail. While Potgieter et al. implemented a Monte Carlo EM algorithm to estimate the model parameters, this study adopted a Bayesian approach similar to van der Linden (2007).

Accuracy Component of the Model

In order to model the accuracy of the reading at the passage level, a binomial count factor model was employed by Potgieter et al. (2017). Thus, it was assumed that the number of correctly read words out of n attempted words in a passage followed a binomial distribution. Let U_{ij} denote the number of words read correctly in passage i by person j . We assume that U_{ij} has a binomial distribution, $U_{ij} \sim B(n_i, p_{ij})$, where n_i is the total number of attempted words in passage i , and p_{ij} is the probability of reading each word in passage i correctly by person j . Potgieter et al. (2017) originally modeled the binomial success probability as a probit model, similar to the parameterization of the two-parameter normal ogive IRT model. The current study modified this binomial success probability similar to the parameterization of the two-parameter logistic IRT model. As a result,

$$p_{ij} = \frac{\exp[a_i(\theta_j - b_i)]}{1 + \exp[a_i(\theta_j - b_i)]}, \quad (1)$$

where a_i and b_i are the discrimination and difficulty parameters, respectively, for passage i , and θ_j is the latent trait parameter describing the reading accuracy of person j . According to this parameterization, the probability of reading words correctly

for a difficult passage will be low, which in turn will result in low number of words read correctly for a given level of θ_j . Conversely, for an easier passage, the number of words read correctly will be higher for the same level of θ_j . The discrimination parameter a_i represents the strength of the association between the probability of reading words correctly for passage i and the latent trait of accuracy.

The choice of the binomial model might seem restrictive; however, Potgieter et al. (2017) argue that it is a reasonable choice for the ORF assessment data for several reasons. First, while it is possible to consider data at the word level, treating passages as units of measurement more closely aligns with the practical application of ORF assessment in the classroom. Also, as demonstrated in Potgieter et al., the number of words read correctly is recovered well with the binomial count approach. Second, there are less restrictive alternative model choices when treating a sentence or passage as a unit of measurement, such as polytomous IRT models. With this approach, however, there will be as many scoring categories as the number of words in each sentence or passage, and it will be rare to observe all possible scoring categories. Therefore, it was determined that such an approach would not be practically appropriate for ORF assessments.

Speed Component of the Model

Time data are positive values and typically exhibit skewness. Therefore, as per van der Linden (2007), the log-normal distribution is a natural and convenient choice for modeling the speed component. Accordingly, Potgieter et al. (2017) used this distribution for reading speed in their model. Let T_{ij} be the time (in seconds) taken to read passage i by person j . Following the log-normal sampling distribution, the natural logarithm of the time variates are assumed to be distributed normally,

$$\ln(T_{ij}) \sim N(\mu_{ij}, \sigma_i^2), \quad (2)$$

where μ_{ij} and σ_i^2 are the mean and variance of the distribution. The distribution function of the time variates is

$$f(t_{ij}; \tau_j, \alpha_i, \beta_i) = \frac{\alpha_i}{t_{ij}\sqrt{2\pi}} \exp \left\{ -\frac{\alpha_i^2}{2} [\ln(t_{ij}) - (\beta_i - \tau_j)]^2 \right\} \quad (3)$$

as described in van der Linden (2007). Here, the mean and variance of this distribution are

$$\mu_{ij} = \beta_i - \tau_j \quad (4)$$

and

$$\sigma_i^2 = \frac{1}{\alpha_i^2}, \quad (5)$$

where τ_j is the latent speed ability for person j , and β_i and α_i are the time intensity (analogous to the time difficulty) and time discrimination parameters for passage i , respectively. Just like the accuracy discrimination parameter, the time discrimination parameter represents the strength of the association between the time to read the passage and the speed ability τ_j . The time intensity parameter β_i indicates a difficulty of the passage regarding reading duration. A larger value of β_i indicates that it will take more time to read the passage given the same latent speed ability τ_j . However, in this formulation, the β_i would be positively associated with the length of passage as it takes more time to read longer passages on average. For this reason, β_i would not be directly comparable between passages of different lengths. Therefore, this study rescaled reading time data to be *per 10 words* ($T_{0ij} = 10 \times T_{ij}/n_i$), where n_i is the number of words in passage i . This way, the time intensity parameter with the rescaled reading time (β_{0i}) would be directly comparable between passages when their lengths are different. Note β_i and β_{0i} are directly related such that $\beta_i = \beta_{0i} + \ln(n_i/10)$.

Relations Between Accuracy and Speed Parameters

Parameters in the accuracy and speed models are jointly modeled and estimated. Specifically, it is assumed that the person-specific accuracy parameter θ_j and speed parameter τ_j are correlated. It is assumed that θ and τ are from a bivariate normal distribution. For identifiability, it is further assumed that the marginal means are both 0, and that the variance of θ is 1.0. The variance of τ and the covariance between θ and τ are treated as part of the model and need to be estimated. The passage parameters are also assumed to be normally distributed with means ($\mu_a, \mu_b, \mu_\alpha, \mu_\beta$) and variances ($\sigma_a^2, \sigma_b^2, \sigma_\alpha^2, \sigma_\beta^2$), and are further assumed to be independent of one another.

Proposed Model-Based Fluency Parameter

This current study extends the work of Potgieter et al. (2017) and proposes a model-based fluency parameter that can be calculated for person j as a function of both the person-specific parameters (θ_j, τ_j) and the passage-specific parameters ($a_i, b_i, \alpha_i, \beta_i$) for $i = 1, \dots, k$ in the above model. In traditional ORF assessments, the observed WCPM score is computed as the observed total number of words read correctly (i.e., accuracy) divided by the observed total reading time in seconds (i.e., speed) and further multiplied by a constant 60. In other words, WCPM is a rate of accurate reading per one unit of 60 seconds. To derive a model-based fluency parameter, the same logic is applied, and thus, the model-based fluency measure is referred to here as the model-based WCPM. Specifically, the model-based WCPM f_j for person j is obtained by dividing the expected value of the total number of words read correctly $E[U_j]$ by the expected value of the total reading time $E[T_j]$, and further multiplied by a constant 60, such that

$$f_j = \frac{E[U_j]}{E[T_j]} \times 60. \quad (6)$$

Here $U_j = \sum_{i=1}^k U_{ij}$ so that $E[U_j] = \sum_{i=1}^k n_i p_{ij}$, where n_i is the number of words in the i th passage, and p_{ij} is the probability of reading each word correctly in the i th passage by person j obtained by Equation (1). Therefore,

$$E[U_j] = \sum_{i=1}^k \frac{n_i \times \exp[a_i(\theta_j - b_i)]}{1 + \exp[a_i(\theta_j - b_i)]}. \quad (7)$$

Similarly, $T_j = \sum_{i=1}^k T_{ij}$ so that $E[T_j] = \sum_{i=1}^k \exp(\mu_{ij} + \frac{1}{2}\sigma_i^2)$, where μ_{ij} is the reading time (in natural logarithm scale) for the i th passage by person j obtained by Equation (4). The exponential transformation should be applied to obtain the expected time on the original scale of reading time. Furthermore, $\mu_{ij} = \beta_i - \tau_j = \beta_{0i} + \log(n_i/10) - \tau_j$, where β_{0i} is the time intensity parameter based on the rescaled reading time. Therefore,

$$E[T_j] = \sum_{i=1}^k \exp\left[\beta_{0i} + \log\left(\frac{n_i}{10}\right) - \tau_j + \frac{1}{2\alpha_i}\right]. \quad (8)$$

As a result, the model-based WCPM score is obtained as a function of passage and person parameters.

Demonstration With Real Data

Data and Model Estimation

For demonstration purposes, a subset of data collected by the larger project to develop an improved ORF assessment system (Nese et al., 2015) was analyzed. The original sample was collected in the 2016-2017 school year in two schools in a Pacific Northwest state of the United States—a total of 1,021 students in Grades 2 to 4 who read up to 10 (of 151) passages. Among the 151 passages, four Grade 4 passages were selected to demonstrate analyses for a set of passages with a total number of words across passages comparable to total words in traditional ORF assessment (approximately 250 words). The total number of words for the selected four passages was 260, where each of the two medium-length passages had 47 words, and the two long passages had 80 and 86 words, respectively. Fourth-grade students who read all selected passages were included in the data for this demonstration. As a result, 58 students were extracted from the original data. Since a sufficient time (90 seconds) was given to read each passage, all sample students read all passages in their entirety. Based on the selected four passages, observed WCPM scores were computed for each student ($M = 110.37$, $SD = 35.50$, $\min = 43.04$, and $\max = 182.94$).

The data were analyzed using JAGS software (Plummer, 2015) that implements a Bayesian Markov chain Monte Carlo (MCMC) algorithm with Gibbs sampling. The

proposed model-based WCPM parameter was embedded in the model estimation procedure using the Bayesian MCMC method such that the posterior distribution of this parameter would be obtained for each individual. As a result, the mean of the posterior distribution was obtained as the point estimate of the model-based WCPM. In addition, the standard deviation of the posterior distribution was obtained as the conditional standard error of measurement (CSEM), and the 95% highest density interval (HDI) of the posterior distribution was obtained as an interval estimate of the model-based WCPM. Noninformative priors were used for all model parameters. Difficulty parameters were assigned normal priors with zero means and large variances (i.e., low precisions) as $b_i \sim N(0, 100)$ and $\beta_i \sim N(0, 100)$. Discrimination parameters were assigned zero-truncated normal priors also with zero means and large variances as $a_i \sim N(0, 100) T(0, \cdot)$ and $\alpha_i \sim N(0, 100) T(0, \cdot)$. The conditional precision of the speed parameter was assigned a Gamma distribution as $\sigma_\tau^{-2} | \theta_j \sim \text{Gamma}(0.01, 0.01)$. Finally, a normal prior with zero mean and large variance was assigned to the covariance between speed and accuracy parameters as $\sigma_{\theta\tau} \sim N(0, 100)$.

In order to ensure proper mixing and convergence of the MCMC procedure, some preliminary analyses were conducted. In these analyses, it was observed that MCMC chains for the difficulty and discrimination parameters of the accuracy part of the model displayed high autocorrelations compared with other model parameters. Thus, it was decided to employ longer chains to ensure sufficiently high effective sample sizes (ESS) for the posterior distributions. Note that we did not employ thinning in order to prevent a loss of precision as suggested by Link and Eaton (2012). In the end, it was decided to run 200,000 iterations after a 10,000 burn-in period in each of three chains, resulting in a total of 600,000 posterior draws for each parameter. As a result, the smallest ESS was 2,521 for one of the speed difficulty parameters, and the largest ESS was 64,134 for one of model-based WCPM parameters. Convergence of the MCMC chains was also evaluated using the Brooks–Gelman–Rubin (BGR; Brooks & Gelman, 1998) statistic. It was confirmed that the BGR values were lower than 1.1 for all model parameters, which was considered to be a good indication of convergence according to Brooks and Gelman. The analysis took approximately 20 minutes using a PC that had a 3.00 GHz Core-i5 CPU with 8 GB memory. The JAGS syntax used for the analysis is provided in Appendix A.

Results

Person-level hyperparameters were estimated to be $\sigma_\tau^2 = 0.09$ (posterior standard deviation = 0.02) and $\sigma_{\theta\tau} = 0.12$ (posterior standard deviation = 0.04). The $\sigma_{\theta\tau}$ on the correlation scale was $\rho_{\theta\tau} = 0.41$ (posterior standard deviation = 0.12), indicating that fast readers in the sample had a tendency to read more accurately. Furthermore, recall that σ_θ^2 was fixed to 1.0 for the purpose of model identification. Estimated passage parameters are reported in Table 1. In terms of accuracy, the easiest passage was Passage 2 ($b = -4.66$), while the most difficult passage was Passage 3 ($b = -2.60$). In terms of speed, Passage 2 was the easiest ($\beta_0 = 3.17$), while Passage 3 was the

most difficult ($\beta_0 = 3.83$). Here, more difficult passages in terms of speed are more time intensive. In other words, it would take more time to read a standard unit (10 words) of the passage, given the same level of speed ability (τ_j).

The estimated model-based WCPM, CSEM, lower and upper bounds of the 95% HDI, and the range of the 95% HDI for the 20 lowest observed WCPMs are presented in Table 2. The same information for the 20 highest observed WCPMs are summarized in Table 3. These tables provide detailed information for the comparison of observed and model-based WCPM scores. It is important to note that the rank order of the observed and model-based WCPM values can change for some of the observations. For example, a student can have a higher model-based WCPM score than another student who has a higher observed WCPM score. However, the results display reasonable coherence between the model-based and observed quantities. The correlation between the observed and model-based WCPMs was nearly perfect at .99. The mean of the difference between observed and model-based WCPMs was $M = -1.03$ ($SD = 6.51$, $\min = -16.73$, and $\max = 16.85$). On the absolute value scale, the mean of the difference was $M = 5.12$ ($SD = 4.10$, $\min = 0.14$, and $\max = 16.85$).

Note that what is evaluated here is the consistency between model-based WCPM and observed WCPM scores, rather than the quality of the model-based WCPM scores. The perfect consistency between the estimated model-based WCPM and the observed WCPM scores would not necessarily indicate a perfect quality of the model-based WCPM scores, because observed WCPMs are also sampled quantities just like model-based WCPMs. For this reason, if an evaluation of the quality of model-based WCPM scores is desired, the model-based scores should be compared with their population counterparts rather than with observed WCPM values. Such an evaluation will be presented in the “Simulation Study” section of this article.

Regarding CSEM, higher model-based WCPM scores were associated with larger CSEM ($r=0.99$), which is consistent with previous research (Stoolmiller et al., 2013). Also, the mean of the CSEM was $M = 8.00$ ($\min = 3.42$, $\max = 12.93$, and $SD = 2.22$). The traditional standard error of measurement (SEM) is analogous to the average of CSEM across a given sample. This mean CSEM is lower than the results from a recent reliability study on the traditional ORF passages from DIBELS (Amplify Education, Inc. & University of Oregon, 2019), where the SEM for Grade

Table 1. Estimated Passage Parameters for Real Data Analysis.

Passage	Word counts	a	b	α	β_0
1	47	0.91	-3.20	6.75	3.23
2	47	0.55	-4.66	9.32	3.17
3	80	0.96	-2.60	5.26	3.83
4	86	0.97	-2.99	6.40	3.68

Note. Third and fourth columns are discrimination and difficulty parameters of the accuracy model. Fifth and sixth columns are discrimination and difficulty parameters of the speed model.

Table 2. Estimated Model-Based Words Correct Per Minute (WCPM) for 20 Lowest Observed WCPM values.

Observed WCPM	Model-based WCPM	CSEM	LB	UB	Range
43.04	42.60	3.42	35.91	49.25	13.35
52.70	50.97	4.17	42.81	59.08	16.27
59.60	57.75	4.38	49.32	66.42	17.10
61.35	57.59	4.59	48.72	66.65	17.93
62.33	67.26	4.96	57.61	77.04	19.44
62.66	69.40	5.29	59.44	80.19	20.74
65.29	74.10	5.61	62.84	84.87	22.02
67.67	68.89	5.06	59.15	78.86	19.71
69.23	75.92	5.84	64.63	87.48	22.85
73.34	82.38	6.42	69.79	94.83	25.03
73.71	84.39	6.59	71.53	97.35	25.83
74.15	73.53	5.64	62.62	84.62	22.00
76.21	76.35	5.59	65.51	87.34	21.83
80.16	81.45	5.97	69.79	93.10	23.30
80.70	83.69	5.92	72.11	95.24	23.13
81.91	81.68	6.23	69.67	94.10	24.43
83.76	84.82	6.40	72.24	97.30	25.06
85.12	94.46	7.01	80.63	108.00	27.37
86.10	87.39	6.33	75.09	99.78	24.70
90.70	94.15	6.92	80.67	107.67	27.00

Note. CSEM is the conditional standard error of measurement. LB and UB are the lower and upper bounds of the 95% highest density interval (HDI), respectively. Range is the width of the 95% HDI.

4 ranged from 9.63 to 12.86. Furthermore, a majority (74.14%) of the sample observations had CSEM lower than 9.63. Research generally indicates that the values of the SEM of traditional curriculum-based measurement of ORF (or CBM-R) measures have been reported to range from 5 to 20 WCPM (e.g., Christ & Silberglitt, 2007; Poncy et al., 2005), and although ORF data with SEM = 5 have been anecdotally described as “very good,” a more realistic range according to published ORF reports is 8 to 10 WCPM.

Simulation Study

Simulation Design

We conducted a Monte Carlo simulation study in order to examine the quality of the model-based WCPM estimates under conditions with varying sample sizes and different numbers of passages. We assumed six passages with lengths of 47, 47, 48, 80, 86, and 86 words per passage. Population passage parameter values used for the simulation study are provided in Table 4. Eighteen simulation conditions were created based on (a) sample sizes (50, 250, or 500), and (b) the numbers of passages (2, 3, 4, 5, or 6)

Table 3. Estimated Model-Based Words Correct Per Minute (WCPM) for 20 Highest Observed WCPM values.

Observed WCPM	Model-based WCPM	CSEM	LB	UB	Range
127.61	125.49	8.95	107.80	142.92	35.12
127.61	116.40	8.85	99.32	133.91	34.59
128.60	125.83	8.93	108.80	143.66	34.86
131.44	129.36	9.22	112.13	148.00	35.87
132.22	128.64	9.11	110.95	146.70	35.75
136.90	130.39	9.24	112.89	149.04	36.15
145.84	138.72	9.88	119.28	157.92	38.64
147.17	144.26	10.79	123.06	165.31	42.25
147.54	143.63	10.48	123.65	164.63	40.98
148.09	138.88	10.10	119.48	158.97	39.49
150.56	154.85	11.21	133.11	176.98	43.87
153.09	140.19	10.00	121.12	160.36	39.23
153.58	151.68	10.74	130.87	173.11	42.24
155.14	149.99	10.65	128.78	170.40	41.62
158.94	147.65	10.57	127.08	168.63	41.55
160.99	159.02	11.16	137.97	181.44	43.47
168.11	161.96	11.48	139.51	184.64	45.13
175.05	158.33	11.66	135.77	181.11	45.34
177.19	167.70	11.93	144.76	191.52	46.76
182.94	178.51	12.93	153.89	204.56	50.67

Note. CSEM is the conditional standard error of measurement. LB and UB are the lower and upper bounds of the 95% highest density interval (HDI), respectively. Range is the width of the 95% HDI.

Table 4. Passage Parameters Used in Simulation Study.

Passage	Word counts	a	b	α	β_0
1	47	0.462	-1.370	3.008	-3.362
2	47	0.532	-1.383	4.799	-4.816
3	48	0.526	-1.336	4.556	-4.366
4	80	0.579	-1.170	4.536	-4.022
5	86	0.640	-1.343	5.289	-5.538
6	86	0.610	-1.261	5.456	-4.462

Note. Time intensity parameter values (the last column) are based on reading time in seconds per 10 words.

with two variations of the three-passage sets, for which the total numbers of words were different. Accordingly, the six passage sets are referred to as (2, 3a, 3b, 4, 5, 6). Set 2 consisted of Passages 1 and 4; Set 3a consisted of Passages 1, 2, and 4; Set 3b consisted of Passages 1, 4, and 5; Set 4 consisted of Passages 1, 2, 4, and 5; Set 5 consisted of Passages 1, 2, 3, 4, and 5; and Set 6 consisted of Passages 1, 2, 3, 4, 5, and 6.

As a result, the total numbers of words were 127, 174, 213, 260, 308, and 394 for the six passage sets (Sets 2, 3a, 3b, 4, 5, and 6), respectively.

Population accuracy and speed parameter values (θ_j and τ_j) were generated for each of the three sample size conditions by using person-level hyperparameters $\sigma_\tau^2 = .124$ and $\sigma_{\theta\tau} = .151$. Then, observed time and count data were randomly generated from the log-normal and binomial distributions by using the related model equations in the preceding sections. The population WCPM values were also obtained based on the population θ_j and τ_j values and population passage parameter values using Equations (6), (7), and (8).

Fifty sets of observed time and count data were randomly generated and analyzed for each of the 18 simulation conditions by JAGS software with syntax similar to the real data analysis in the previous section. The quality of the estimated model-based WCPM scores was evaluated by the absolute bias, which compared them with their population counterparts. Also, the standard errors were calculated as the standard deviation of the estimated model-based WCPM scores from the 50 replications. Since the quality of the estimated model-based WCPM scores depended on the population WCPM values, absolute bias and standard errors were averaged conditioned on the population WCPM values. Specifically, average absolute bias (AAB) and average standard error (ASE) values were obtained for four predefined ranges of population WCPM values; $WCPM < 50$, $50 \leq WCPM < 100$, $100 \leq WCPM < 150$, and $WCPM \geq 150$.

We also conducted a series of sensitivity analyses to evaluate the effect of prior distributions with different levels of information on the estimation of the model-based WCPM scores. The results revealed that the recovery of the model-based WCPM scores was not affected by the choice of priors. Thus, we concluded that the use of noninformative priors was supported, and the same set of noninformative priors as with the real data analysis was used in the simulation study. Detailed results of the sensitivity analyses are provided in Appendix B.

Simulation Results

AABs and ASEs are summarized in Figures 1 and 2, respectively. First, both AAB and ASE decreased as the number of the passages increased. Also, both AAB and ASE were higher for higher population WCPM groups. In the real data analysis in the previous section, we observed that higher WCPMs were associated with higher CSEMs, and this tendency was confirmed in the simulation study. The difference between the two 3-passage conditions (3a and 3b) was demonstrated both on AAB and ASE, except for $N = 250$ and $N = 500$ conditions for $WCPM < 50$. This result demonstrated that the number of words did have some effect on the quality of the estimated model-based WCPM, in addition to the number of passages. However, the difference between 3a and 3b conditions appeared to be smaller than the difference between conditions with different numbers of passages.

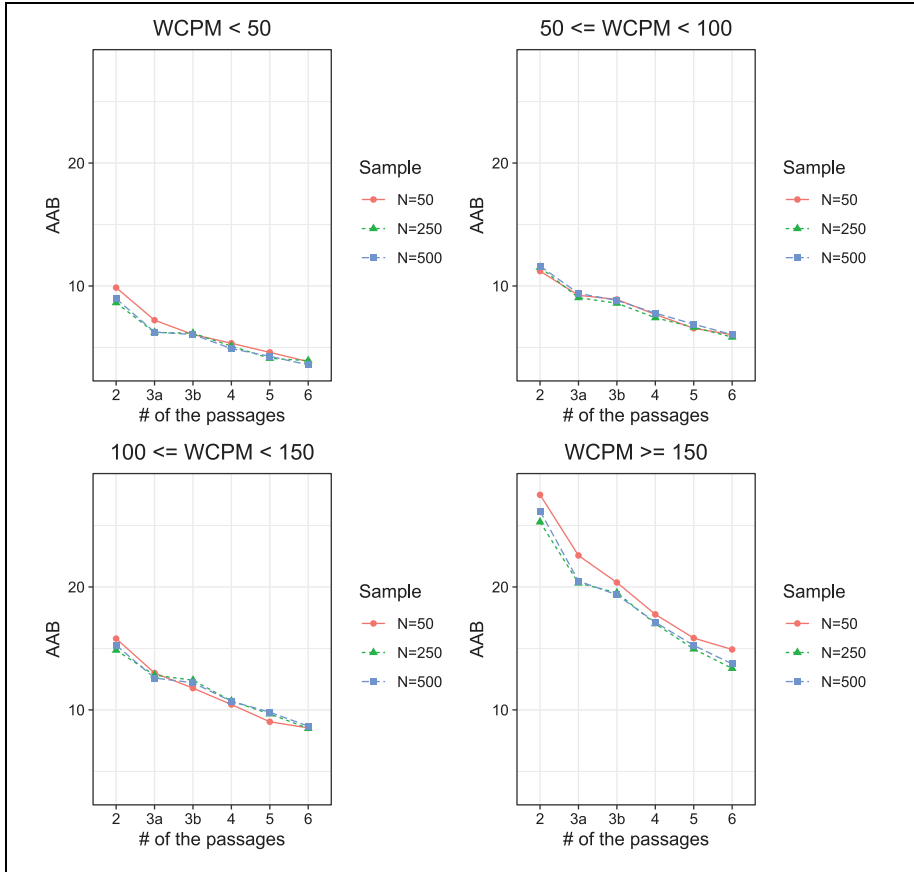


Figure 1. Average absolute bias (AAB) of model-based words correct per minute (WCPM) estimates.

Second, the effect of sample size was not as clear as the effect of the number of passages. Increased sample sizes did not always result in substantial improvements in AAB and ASE. While the smallest sample size ($N = 50$) had substantially larger AAB and ASE values for only some of the passage conditions, in some instances it demonstrated equivalent quality of the estimated model-based WCPM values obtained for larger sample sizes. The effect of sample size was more clear for the range of $WCPM \geq 150$, where AAB and ASE values for $N = 250$ and $N = 500$ were generally lower than the sample size of $N = 50$. Overall, the results revealed that the number of passages were more important for the quality of the estimated model-based WCPM scores than the sample size. Also, we conclude that the quality of the model-based WCPM scores is quite reasonable even with a small sample size like $N = 50$ as long as data are collected for a sufficient number of passages.

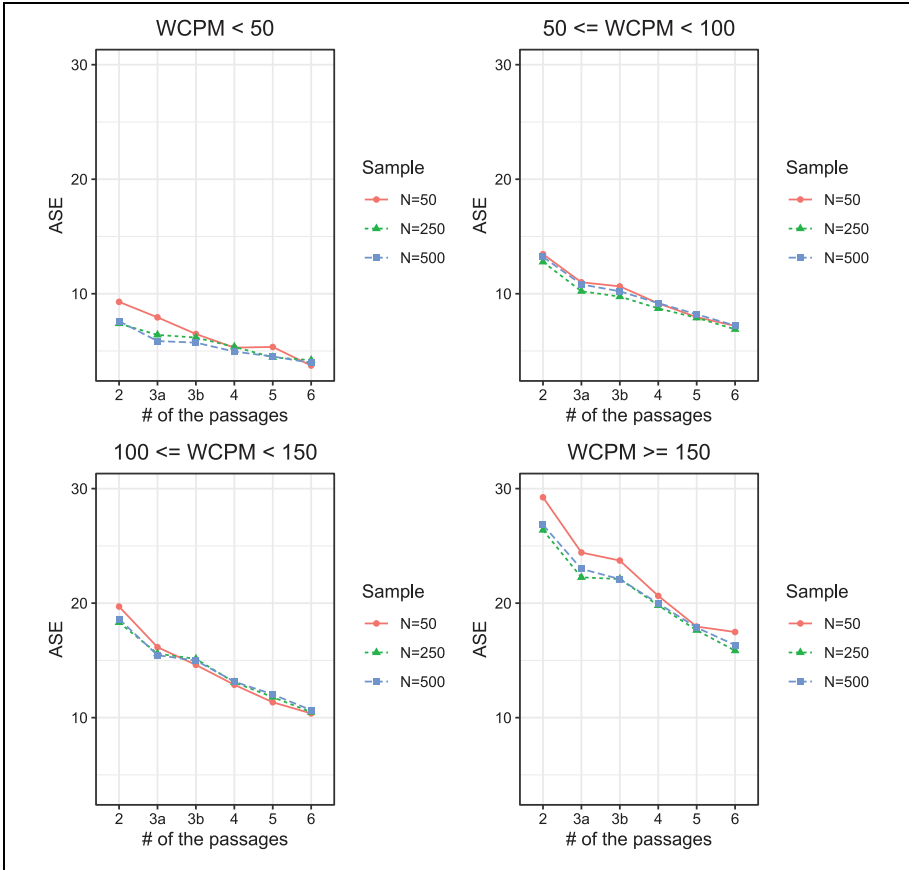


Figure 2. Average standard errors (ASE) of model-based words correct per minute (WCPM) estimates.

Discussion

This study proposed a new model-based WCPM score as a measure of ORF. The procedure was demonstrated with a real data set of 58 students who read four passages with lengths ranging from 47 to 86 words for a total of 260 words read, approximately equivalent to a traditional ORF assessment. The results of the real data analysis demonstrated reasonable coherence between observed and model-based WCPM scores. Also, it was observed that the mean CSEM for the model-based WCPM scores was lower than SEMs reported for the traditional ORF assessment. This suggests that SEM estimates comparable to or better than those of traditional ORF assessments may be achievable using a small number of passages, and accordingly through shorter assessment administration.

The results of the simulation study demonstrated that the quality of the model-based WCPM estimates improved when estimated with a larger number of passages. Although small improvements were also observed with larger sample sizes, the effect of sample size was not as clear as the effect of the number of passages, and good quality fluency estimates were achieved even with a sample size of 50. In an applied setting, the optimal number of passages to administer may be determined by evaluating the CSEMs compared with a desired value of SEM. For example, one may wish to see a majority of observations (e.g., more than 50%) in the sample have CSEM smaller than a target level of SEM, such as 8.0. One can then conduct a Monte-Carlo simulation to evaluate the distribution of CSEMs to find the number of passages meeting these criteria.

This study demonstrated the procedure to estimate model-based WCPM scores and assessed their quality. However, the utility of the model-based WCPM scores is important beyond the context of the current study where all students read all passages. Specifically, the model-based WCPM easily extends to scenarios with missing data. In practical settings, it is likely to have more passages than the number of test occasions for test security and exposure control purposes. This results in each student only reading a small subset of passages from a larger passage pool with the specific subset varying both across students and measurement occasions. In such a circumstance, there will be a large proportion of missing data in the combined data set. Even so, the Bayesian methodology utilized in this study can still estimate passage parameters when missing data are present. This enables one to conduct passage-parameter equating by carefully designing and assembling assessment forms that contain common passages between forms, and concurrently estimate passage parameters for all passages. This is the same idea as test equating using the concurrent equating method based on a common-item nonequivalent group design with an IRT model where test item parameters are equated in the same scale. Just like conventional test items, ORF assessment passages can also be calibrated to equate passage parameters by various equating designs and methods, such as common-passage nonequivalent group design with concurrent passage calibrations by utilizing a model such as the one utilized in this study.

Once the passage parameters are equated in the same scale, estimated person-specific accuracy and speed parameters (θ_j and τ_j) will be comparable across students, regardless of which subset of passages the student read. Then, model-based WCPMs can be computed for all measurement occasions and all students using one selected set of passages as a reference. This selected subset of passages does not have to be read by all students; however, the passages have to be from an equated passage pool and have to be common to all measurement occasions and all students. This will result in model-based WCPM scores that are comparable to each other no matter what set of passages each student read. As such, score comparability within and between students will not be an issue for reporting WCPM scores without any effort for post-equating (e.g., Stoolmiller et al., 2013), which is typically required by traditional equating approaches.

Practical Implications

The implications of this study extend beyond the field of measurement to the classroom. Research has shown that variance in “difficulty” across ORF passages leads to construct irrelevant variance and results in less accurate measures of growth (Francis et al., 2008). The model-based WCPM offers educators scores that are comparable regardless of the passages read, and is a vast improvement on the common practice of using readability estimates (e.g., Flesch-Kincaid) to equate passages (Betts et al., 2009; Poncy et al., 2005). In addition, students are generally administered ORF assessments at their grade level, even if that does not match their instructional level, partly because it is unfeasible to draw inferences about their reading proficiency given off-grade-level assessments, and partly because universal screening measures are to be standardized. The approach taken here can mitigate the constraint of grade-level assessments by applying a common model-based WCPM scale across grades, providing an advantage for growth analyses of student ORF across grades. Perhaps most important, the reduction of the CSEM, particularly for low-performing students, improves the reliability of ORF scores and yields scores sensitive to instructional change. The model-based WCPM scores are arguably better suited for measuring ORF, both for screening and progress monitoring, as a more precise score will lead to more accurate instructional decisions. They also have the important distinction of being on the same metric as traditional ORF scores—WCPM. This makes the scale scores immediately usable for teachers and reading specialists who are familiar with the WCPM expectations for students at specific times in specific grades along the reading continuum. As ORF assessments are used by researchers, teachers, and administrators, the model-based WCPM scores can increase the reliability and validity of the decisions made from scores, yielding better identification of students in need of reading interventions, and better evaluation of the results of those interventions.

Limitations and Future Research

Going forward, some extensions are desired to advance the proposed model-based WCPM approach. First, it is important that the computation algorithm is easily available to estimate model-based WCPM scores based on calibrated passage parameters. As mentioned earlier, a typical practical scenario to estimate model-based WCPM scores would be when one has a pool of many passages with equated passage parameters. In fact, although it is not presented in this article, the JAGS syntax has been modified (actually simplified) to a case where we have known passage parameters. Second, related to the first point, it is desired to have a faster algorithm to compute model-based WCPM scores, such as by maximum likelihood and/or faster Bayesian algorithms. As mentioned previously, the JAGS syntax to estimate passage parameters for four passages and 58 model-based WCPM scores took about 20 minutes. When the JAGS syntax was simplified to estimate only 58 model-based WCPM scores by assuming passage parameters are known, the computation time was

substantially improved to about 8 minutes. However, this is still not fast enough to handle a larger data set. For example, it would take approximately 2.5 hours if we had comparable reading data for 1,000 students. Last, future research is needed to extend the study model where sentences are the unit of analysis. This will be potentially helpful to improve the quality of model parameter estimates, and as a result, the quality of the estimated model-based WCPM scores. Although the current model can accommodate accuracy and time data at the sentence level, there would be a need to incorporate dependency between sentences within the same passage if the data included more than one passage. This is a similar issue as a testlet, such as a set of reading comprehension items associated with the same reading passage.

Appendix A

JAGS Syntax for Real Data Analysis

The following syntax assumes J students and I passages. Data are (a) $res = J \times I$ matrix of the number of words correctly read, (b) $nw = J \times I$ matrix of the number of words in the passages, and (c) $tim = J \times I$ matrix of the reading time per 10 words in the natural logarithm scale.

```

model{
  # J students and I passages
  for (j in 1:J) {
    for (i in 1:I) {
      res[j,i] ~ dbin(p[j,i], nw[j,i])
      cnt_ex[j,i] <- p[j,i] * nw[j,i]
      logit(p[j,i]) <- a[i] * (theta[j] - b[i])
      tim[j,i] ~ dnorm(mu[j,i], prec.t[i])
      mu[j,i] <- beta[i] - tau[j]
      tim_ex[j,i] <- exp(mu[j,i] + log(nw[j,i]/10)
        + 0.5 * 1/ (pow(alpha[i], 2)))
    }
    theta[j] ~ dnorm(0, 1)
    tau[j] ~ dnorm(mtau[j], ptau)
    mtau[j] <- cvr * theta[j]
    exp_cnt[j] <- sum(cnt_ex[j,])
    exp_min[j] <- sum(tim_ex[j,])
    orf[j] <- exp_cnt[j]/exp_min[j] * 60 # Model-based WCPM
  }
  # Priors for passage parameters
  for(i in 1:I) {
    prec.t[i] <- pow(alpha[i], 2)
    alpha[i] ~ dnorm(0, 0.01) I(0,)
    beta[i] ~ dnorm(0, 0.01)
    a[i] ~ dnorm(0, 0.01) I(0,)
  }
}

```

```

    b[i] ~ dnorm(0, 0.01)
  }
  # Priors for person parameters
  ptau ~ dgamma(0.01, 0.01)
  vtau <- 1/ptau
  tau.var <- vtau + (pow(cvr, 2))
  cvr ~ dnorm(0, 0.01)
  crl <- cvr/sqrt(tau.var)
}

```

Appendix B

Sensitivity Analyses

A series of sensitivity analyses were conducted to examine the effect of using priors with different levels of information on the estimation of the model-based WCPMs. We focused on a specific condition with $N = 250$ and six passages. In addition to the same noninformative priors used in real data analysis, a set of mildly and highly informative priors were considered for passage parameters and person hyperparameters. For the passage parameters ($a_i, b_i, \alpha_i, \beta_i$) and the covariance between the speed and accuracy parameters ($\sigma_{\theta\tau}$), the variance of the noninformative normal priors were 10 and 1 for mildly and highly informative priors, respectively. For the conditional precision of the speed parameter ($\sigma_{\tau}^{-2}|\theta_j$), the shape and rate parameters were set to 0.1 for mildly informative priors, and to 1.0 for highly informative priors.

Five specifications of priors were examined: (a) priors for both passage and person parameters were noninformative (vague); (b) priors for passage parameters were mildly informative, while priors for person parameters were noninformative; (c) priors for passage parameters were highly informative, while priors for person parameters were noninformative; (d) priors for person parameters were mildly informative, while priors for passage parameters were noninformative; and (e) priors for person parameters were highly informative, while priors for passage parameters were noninformative.

Fifty data sets were generated and the model-based WCPMs were estimated with the five different specifications of the priors. The recovery of the model-based WCPM scores were examined by the average standard error (ASE) and average absolute bias (AAB) values for the four ranges of population WCPM scores ($WCPM < 50$, $50 \leq WCPM < 100$, $100 \leq WCPM < 150$, and $WCPM \geq 150$). Results of the sensitivity analyses revealed that the recovery of the model-based WCPM scores was not affected by the different priors. The AAB and ASE values for the five different prior specifications were close to each other for all four ranges of WCPM scores. Results are graphically presented in Figures B1 and B2.

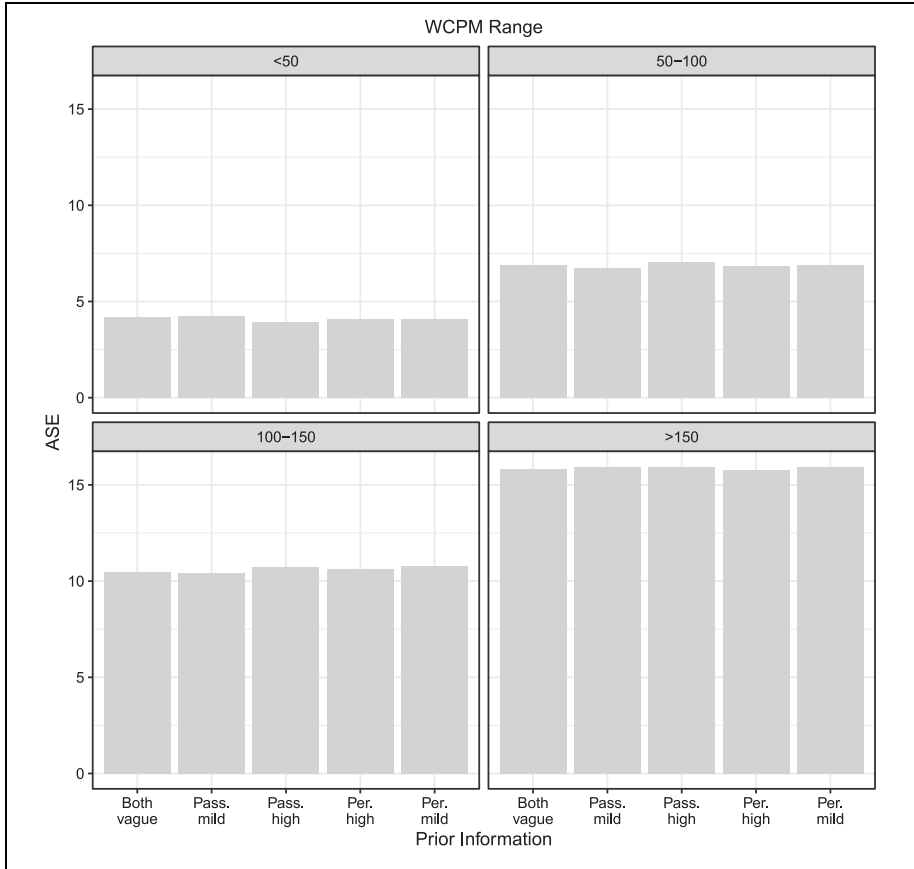


Figure B1. Average standard errors (ASE) for different specifications of priors.

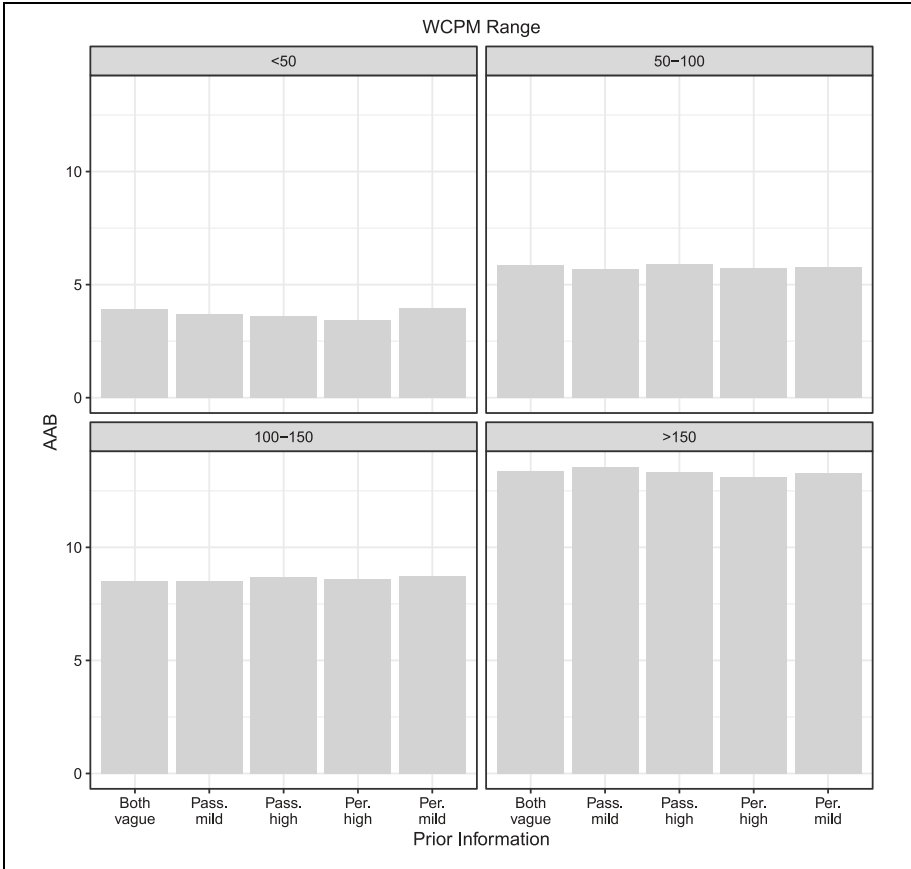


Figure B2. Average absolute bias (AAB) for different specifications of priors.


Declaration of Conflicting Interests

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

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: The research reported here was partially 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.

ORCID iD

Akihito Kamata  <https://orcid.org/0000-0001-9570-1464>

References

- Amplify Education, Inc., & University of Oregon. (2019). *DIBELS 8th editions: Administration and scoring guide*. https://dibels.uoregon.edu/docs/materials/d8/dibels_8_admin_and_scoring_guide_09_2019.pdf
- 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. <https://doi.org/10.1016/j.jsp.2008.09.001>
- Brooks, S. P., & Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics, 7*(4), 434-455. <https://doi.org/10.1080/10618600.1998.10474787>
- Christ, T. J., & Silbergliitt, B. (2007). Estimates of the standard error of measurement for curriculum-based measures of oral reading fluency. *School Psychology Review, 36*(1), 130-146.
- Colorado Department of Education. (n.d.). *Colorado student assessment program*. <http://www.cde.state.co.us/>
- Crawford, L., Tindal, G., & Stieber, S. (2001). Using oral reading rate to predict student performance on statewide achievement tests. *Educational Assessment, 7*(4), 303-323. https://doi.org/10.1207/S15326977EA0704_04
- Deno, S. L. (1985). Curriculum-based measurement: The emerging alternative. *Exceptional Children, 52*(3), 219-232. <https://doi.org/10.1177/001440298505200303>
- Entink, R. H. K., Kuhn, J. T., Hornke, L. F., & Fox, J. P. (2009). Evaluating cognitive theory: A joint modeling approach using responses and response times. *Psychological Methods, 14*(1), 54-75. <https://doi.org/10.1037/a0014877>
- Fox, J. P., Klein Entink, R. H. K., & van der Linden, W. J. (2007). Modeling responses and response times with package cirt. *Journal of Statistical Software, 20*(7), 1-14. <https://doi.org/10.18637/jss.v020.i07>
- Francis, D. J., Santi, K. L., Barr, C., Fletcher, J. M., Varisco, A., & Fooman, B. R. (2008). Form effects on the estimation of students' oral reading fluency using DIBELS. *Journal of School Psychology, 46*(3), 315-342. <https://doi.org/10.1016/j.jsp.2007.06.003>
- Fuchs, L. S. (2004). The past, present, and future of curriculum-based measurement research. *School Psychology Review, 33*(2), 188-192.
- Fuchs, L. S., Fuch, D., Hosp, M. K., & Jenkins, J. R. (2001). Oral reading fluency as an indicator of reading competence: A theoretical, empirical, and historical analysis. *Scientific Studies of Reading, 5*(3), 239-256. https://doi.org/10.1207/S1532799XSSR0503_3
- Fuchs, L. S., Fuch, D., & Maxwell, L. (1988). The validity of informal reading comprehension measures. *Remedial and Special Education, 9*(2), 20-28. <https://doi.org/10.1177/074193258800900206>
- Harcourt Brace. (2003). *Stanford achievement test—tenth edition: Technical data report*.
- Hasbrouck, J., & Tindal, G. (2006). Oral reading fluency norms: A valuable assessment tool for reading teachers. *The Reading Teacher, 59*(7), 636-644. <https://doi.org/10.1598/RT.59.7.3>
- Hasbrouck, J., & Tindal, G. (2017). *An update to compiled ORF norms* (Technical Report No. 1702). University of Oregon.
- Hoover, H. D., Dunbar, S. B., & Frisbie, D. A. (2001). *Iowa tests of basic skills*. Riverside.
- Hudson, R. F., Lane, H. B., & Pullen, P. C. (2005). Reading fluency assessment and instruction: What, why, and how? *The Reading Teacher, 58*(8), 702-714. <https://doi.org/10.1598/RT.58.8.1>

- Jenkins, J. R., Fuchs, L. S., Van Den Broek, P., Espin, C., & Deno, S. L. (2003). Sources of individual differences in reading comprehension and reading fluency. *Journal of Educational Psychology, 95*(4), 719-729. <https://doi.org/10.1037/0022-0663.95.4.719>
- Kim, Y. S., Petscher, Y., Schatschneider, C., & Foorman, B. R. (2010). Does growth in oral reading fluency matter in reading comprehension achievement? *Journal of Educational Psychology, 102*(3), 652-667. <https://doi.org/10.1037/a0019643>
- LaBerge, D., & Samuels, S. J. (1974). Toward a theory of automatic information processing in reading. *Cognitive Psychology, 6*(2), 293-323. [https://doi.org/10.1016/0010-0285\(74\)90015-2](https://doi.org/10.1016/0010-0285(74)90015-2)
- Link, W. A., & Eaton, M. J. (2012). On thinning of chains in MCMC. *Methods in Ecology and Evolution, 3*, 112-115. <https://doi.org/10.1111/j.2041-210X.2011.00131.x>
- Miura Wayman, M., Wallace, T., Wiley, H. I., Tichá, R., & Espin, C. A. (2007). Literature synthesis on curriculum-based measurement in reading. *Journal of Special Education, 41*(2), 85-120. <https://doi.org/10.1177/00224669070410020401>
- National Reading Panel. (2000). *Report of the national reading panel: 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*. National Institute of Child Health and Human Development.
- Nese, J. F.T., Biancarosa, G., Cummings, K., Kennedy, P., Alonzo, J., & Tindal, G. (2013). In search of average growth: Describing within-year oral reading fluency growth across grades 1–8. *Journal of School Psychology, 51*(5), 625-642. <https://doi.org/10.1016/j.jsp.2013.05.006>
- Nese, J. F. T., & Kamata, A. (2019). *Comparing scoring methods and passage length of CBM-r: Convergent and content evidence for automated scoring and shorter passages* [Manuscript submitted for publication].
- Nese, J. F. T., Kamata, A., & Alonzo, J. (2015, July 15-18). *Exploring the evidence of speech recognition and shorter passage length in computerized oral reading fluency*. Paper presented at the 22nd annual meeting of the Society for the Scientific Study of Reading, Kailua-Kona, HI.
- Perfetti, C. A. (1985). *Reading ability*. Oxford University Press.
- Pinnell, G., Pikulski, J., Wixson, K., Campbell, J., Gough, P., & Beatty, A. (1995). *Listening to children read aloud*. US Department of Education, National Center for Education Statistics.
- Plummer, M. (2015). *JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling* (Version 4.0.0). <https://sourceforge.net/projects/mcmc-jags/>
- 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. <https://doi.org/10.1177/073428290502300403>
- Potgieter, C. J., Kamata, A., & Kara, Y. (2017). *An EM algorithm for estimating an oral reading speed and accuracy model*. <https://arxiv.org/abs/1705.10446>
- Ranger, J. (2013). A note on the hierarchical model for responses and response times in tests of van der Linden. *Psychometrika, 78*(3), 538-544. <https://doi.org/10.1007/s11336-013-9324-6>
- Roehrig, A. D., Petscher, Y., Nettles, S. M., Hudson, R. F., & Torgesen, J. K. (2008). Accuracy of the DIBELS oral reading fluency measure for predicting third grade reading comprehension outcomes. *Journal of School Psychology, 46*(3), 343-366. <https://doi.org/10.1016/j.jsp.2007.06.006>

- Shinn, M. R., Good, R. H., Knutson, N., Tilly, W. D., & Collins, V. L. (1992). Curriculum-based measurement or oral reading fluency: A confirmatory analysis of its relation to reading. *School Psychology Review, 21*(3), 459-479. <https://doi.org/10.1080/02796015.1992.12085629>
- Stoolmiller, M., Biancarosa, G., & Fien, H. (2013). Measurement properties of DIBELS oral reading fluency in grade 2: Implications for equating studies. *Assessment for Effective Intervention, 38*(2), 76-90. <https://doi.org/10.1177/1534508412456729>
- Valencia, S., Smith, A., Reece, A., Li, M., Wixson, K., & Newman, H. (2010). Oral reading fluency assessment: Issues of construct, criterion, and consequential validity. *Reading Research Quarterly, 45*(3), 270-291. <https://doi.org/10.1598/RRQ.45.3.1>
- van der Linden, W. J. (2007). A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika, 72*(3), 287-308. <https://doi.org/10.1007/s11336-006-1478-z>
- van der Linden, W. J., Entink, R. H. K., & Fox, J. P. (2010). IRT parameter estimation with response times as collateral information. *Applied Psychological Measurement, 34*(5), 327-347. <https://doi.org/10.1177/0146621609349800>
- Wood, D. E. (2006). Modeling the relationship between oral reading fluency and performance on a statewide reading test. *Educational Assessment, 11*(2), 85-104. https://doi.org/10.1207/s15326977ea1102_1
- Yovanoff, P., Duesbery, L., Alonzo, J., & Tindal, G. (2005). Grade-level invariance of a theoretical causal structure predicting reading comprehension with vocabulary and oral reading fluency. *Educational Measurement: Issues and Practice, 24*(3), 4-12. <https://doi.org/10.1111/j.1745-3992.2005.00014.x>