Equating Oral Reading Fluency Scores: A Model-Based Approach

Yusuf Kara[1], Akihito Kamata[2], Xin Qiao[3], Cornelis J. Potgieter[4], & Joseph F. T. Nese[5]

[1]Southern Methodist University, ORCID ID:0000-0003-0691-0630

[2]Southern Methodist University, ORCID ID: 0000-0001-9570-1464

[3]Southern Methodist University, ORCID ID: 0000-0002-3248-7859

[4]Texas Christian University, ORCID ID: 0000-0002-1995-6817

[5]University of Oregon, ORCID ID: 0000-0002-9878-7395

Paper published online first in Educational and Psychological Measurement on 1/5/2023.

https://doi.org/10.1177/00131644221148122

Author Note

The research reported here was partially supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305D200038 to the Southern Methodist University and 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.

Correspondence concerning this article should be addressed to Yusuf Kara, Center on Research and Evaluation, Southern Methodist University, Dallas, TX 75275. E-mail: ykara@smu.edu

Table of Contents

Abstract

Words read correctly per minute (WCPM) is the reporting score metric in oral reading fluency (ORF) assessments, which is popularly utilized as part of curriculum-based measurements to screen at-risk readers and to monitor progress of students who receive interventions. Just like other types of assessments with multiple forms, equating would be necessary when WCPM scores are obtained from multiple ORF passages to be compared both between and within students. This paper proposes a model-based approach for equating WCPM scores. A simulation study was conducted to evaluate the performance of the model-based equating approach along with some observed-score equating methods with external anchor test design.

Keywords: oral reading fluency, curriculum-based measurement, words read correctly per minute, model-based approach, equating

Equating Oral Reading Fluency Scores: A Model-Based Approach

Oral reading fluency (ORF) has been regarded as an important indicator of overall reading competency and assessed frequently as part of curriculum-based measurements to screen at-risk readers and to monitor progress of students who receive interventions. In a typical administration of ORF assessment, a student is given a grade-level text to read, and the number of words read correctly per minute (WCPM) is computed based on the observed number of correctly read words and the observed reading time. WCPM scores have been the most commonly-used measure in ORF assessments in classrooms, as well as in national-level assessments (e.g., White et al., 2021). Previous research provided empirical evidence on the predictive and concurrent validity of WCPM scores (e.g., Fuchs et al., 2001; Hasbrouck & Tindal, 2006). Despite their prevalent use and practical applications, observed WCPM scores have considerable psychometric limitations such as inaccurate standard errors (Christ & Silberglitt, 2007; Nese et al., 2013) and dependence to specific passages read by students (Betts et al., 2009; Francis et al., 2008), which potentially reduces the reliability and validity of reported scores.

Passage dependence of the observed WCPM scores points to the variability in passage difficulties in ORF assessments. In other words, a student may have different WCPM scores by reading an easy or hard passage at the same grade-level. In a similar vein, two students with the same level of ORF ability may yield different WCPM scores due to reading passages that have different difficulty levels. Also referred to as “passage effects” (Cummings, et al., 2013), this variability in passage difficulties may produce a considerable amount of systematic error, as much as 10% (around 22 WCPM; Chaparro et al., 2018). Such a large magnitude of error in ORF scores can be crucial, especially for students who are identified to be at risk of poor reading

outcomes. Moreover, it would be difficult to distinguish the change of ORF scores due to true student growth versus passage dissimilarity, which would jeopardize the validity of longitudinal ORF assessments (Albano et al., 2018; Albano & Rodriguez, 2012). Thus, test developers of ORF assessments need to correct for varying passage difficulties with appropriate methods for more precise measurement of ORF.

Commonly accepted practices for handling passage effects include evaluating passage difficulties through readability indices (e.g., Flesh-Kincaid) and increasing the number of passages read by the student at each session. However, these strategies can generally be insufficient for establishing passage equivalence (Betts et al., 2009; Cummings et al., 2013; Francis et al., 2008; Stoolmiller et al., 2013). For example, the same passage may yield a different readability index score based on a specific index used, which creates another complexity in handling passage effects in ORF assessments through readability indices (Good & Kaminski, 2002; Yi, 2021). Also, Albano and Rodriguez (2012) highlighted the difficulty of obtaining parallel passages based on such readability indices.

As a more comprehensive approach to account for passage effects, test score equating procedures have been used by some researchers and practitioners. The main motivation for equating is to ensure that WCPM scores from different passages can be used interchangeably. In addition to a horizontal equating of WCPM scores that aims to account for within-grade passage effects, a vertical equating across grade-levels is thought to be an important effort, especially for monitoring progress of young readers over multiple years. Referred to as observed-score equating methods in psychometric literature, researchers mostly have adapted mean, linear, and equipercentile equating methods (Albano & Rodriguez, 2012; Santi et al., 2016; Stoolmiller et al., 2013) to equate observed WCPM scores across passages. These observed-score equating

procedures are sample specific. Therefore, unless one has a population-representative equating sample, the results of equating would not be generalizable to other samples. This will make the use of observed-score equating procedures difficult to justify for the purpose of a pre-equating.

On the other hand, a model-based approach to ORF assessment data has a potential to overcome these shortcomings. An estimation approach of model-based WCPM scores has recently been introduced by Kara et al. (2020) based on a latent-variable psychometric model for ORF assessment data (Potgieter et al., 2017) as part of an effort to develop an improved computer-based ORF assessment system (Nese & Kamata, 2014-2018). Model-based WCPM scores are in the form of expected WCPM values estimated with the latent-variable model that incorporates person and passage parameters. Please note that we explicitly distinguish the traditional WCPM scores from the model-based WCPM scores by referring the former to as observed WCPM scores in this paper.

Strengths of the model-based WCPM scores such as higher reliability and availability of conditional measurement errors have been demonstrated by Nese and Kamata (2021). In addition to having better psychometric characteristics, the model-based WCPM scores can eliminate the need for post-equating. More specifically, the equating procedure by the model-based approach allows one to develop a pool of passages where all passage parameters are calibrated into the same scale by the common-item nonequivalent group (NEG) design. Similar to item response theory (IRT) approach for equating test scores, an equated passage pool allows one to estimate student-level speed and accuracy latent factor scores comparable to each other, no matter which set of passages the student read. As a result, it will be also possible to place the model-based WCPM scores on a common scale.

This study aims to introduce a model-based approach to equate WCPM scores based on Kara et al. (2020) and to provide a numerical demonstration with simulated data. In the numerical demonstrations, we designed a simulation study to compare performance of the proposed model-based approach with traditional observed-score equating methods. The structure of the paper is as follows. We first describe the procedures of the proposed model-based approach for equating WCPM scores, including a brief presentation of the model-based WCPM score estimation. Next, we briefly provide information about two observed-score equating methods, namely, linear and equipercentile equating. In the following sections, we describe the details of the simulation study, present the results, and conclude with a discussion section.

Equating WCPM Scores

Model-Based Estimation of WCPM Scores

Estimation of the model-based WCPM scores has recently been introduced by Kara et al. (2020) based on a latent-variable psychometric model with speed and accuracy components (Potgieter et al., 2017). Note that this model-based approach is different from the latent variable approach demonstrated in Stoolmiller et al. (2013) and Santi et al. (2016), which use observed WCPM scores as the observed indicators of the latent ORF ability in a confirmatory factor model. These factor models can be considered as congeneric models in the classical test theory, which is essentially a model for weighted summed observed scores. Rather, the model for current approach was developed as a joint factor model of speed and accuracy, which is a modification of the hierarchical speed and accuracy model (van der Linden, 2007). In this model, passage-level observed reading times and the number of correctly read words are used as the observed indicators of two separate latent factors, namely, one latent factor for speed and another latent factor for accuracy.

The speed component of the model uses the same log-normal factor model as in van der Linden (2007). Thus, the natural logarithm of t_{ij}, time taken (in seconds) to read passage *i* by person *j,* assumed to have a distribution function as

$$f(t_{ij};\tau_j,\alpha_i,\beta_i) = \frac{\alpha_i}{t_{ij}\sqrt{2\pi}} \exp\left\{-\frac{\alpha_i^2}{2}\left[\ln(t_{ij}) - (\beta_i - \tau_j)\right]^2\right\}, \quad (1)$$

where τ_j is the latent speed ability for person j. β_i and α_i are the time intensity and discrimination parameters for passage i, respectively. Since the magnitude of the time intensity parameter depends on the passage length (i.e., a longer passage takes more time to read), it would be desirable to rescale the time intensity parameter such as in a scale of reading time per 10 words. This rescaled time intensity parameter can be formulated as $\beta_{0i} = \beta_i - \log(n_i/10)$, where n_i is the number of the words that passage *i* contains.

Accuracy component of the model uses a binomial-count factor model. The number of the words read correctly in passage *i* (out of n_i words available) by person *j* is assumed to be drawn from a binomial distribution with a success probability (i.e., reading the word correctly) of p_{ij} per word. Using a logit link function and the same parametrization as the two-parameter IRT model, p_{ij} is further modeled as

$$p_{ij} = \frac{\exp\left[a_i(\theta_j - b_i)\right]}{1 + \exp\left[a_i(\theta_j - b_i)\right]}, \quad (2)$$

where θ_j is the latent accuracy ability of person *j.* a_i and b_i are discrimination and difficulty parameters of passage *i* in terms of reading accuracy.

Similar to true scores estimated based on an IRT model, a model-based WCPM score for person *j* ($\hat{f}_j$) is calculated from the person and passage parameter estimates obtained from the speed and accuracy components defined above. Following the traditional definition that WCPM

is a rate of accurate reading per 60 seconds, $\hat{f}_j$ is calculated as the expected value of the total number of words read correctly $E[U_{.j}]$ divided by the expected value of the total reading time in seconds $E[T_{.j}]$ and further multiplied by a constant of 60 as follows:

$$f_j = \frac{E[U_{.j}]}{E[T_{.j}]} \times 60, \tag{3}$$

where

$$E[U_{.j}] = \sum_{i=1}^{k} n_i p_{ij} \tag{4}$$

in which n_i is the number of words in passage *i*, *k* is the number of passages read by person *j*, and p_{ij} is defined in Eq. 2. Similarly,

$$E[T_{.j}] = \sum_{i=1}^{k} \exp\left(\beta_{0i} + \log\left(\frac{n_i}{10}\right) - \tau_j + \frac{1}{2\alpha_i^2}\right), \tag{5}$$

where τ_j, and α_i are defined in Eq. 1, while β_{0i} is the rescaled time intensity parameter such that $E[T_{.j}]$ is on the original scale of reading time (in seconds). A more detailed derivation of $E[T_{.j}]$ can be found in Kara et al. (2020).

Note that θ and τ must be estimated based on the set of passages that the student read. However, $E[U_{.j}]$ and $E[T_{.j}]$, as well as f_j, can be computed for a set of passages that the student read or did not read. For the latter case, passage parameters should have been calibrated into the same scale as the passages that student read to estimate θ and τ. This is a key idea to equating WCPM scores by the model-based approach, which will be described in more detail in the next section. Estimation of all model parameters including the model-based WCPM scores can be done by adopting a Bayesian approach. Readers are referred to Kara et al. (2020) for more details regarding the parameter estimation and further details of the latent-variable model.

Proposed Equating Method for Model-Based WCPM Scores

Being a latent-variable psychometric model, the measurement model for ORF assessment data by Potgieter et al. (2017) has the same advantages of traditional IRT models in equating studies. In our application of the latent-variable ORF model, passages are analogous to items. As indicated by Kolen and Brennan (2004), IRT models' strengths stem from modeling response data at the item-level unlike the classical test theory, which focuses on test-level data (i.e., observed total scores). For the ORF latent-variable model, time intensity and difficulty parameters, as well as two types of discrimination parameters for each passage, are estimated from the time and accuracy components of the model. At the person level, speed and accuracy ability parameters are estimated first. Then, a model-based WCPM score is ultimately estimated as a measure of the ORF ability. We propose a four-step approach to produce equated model-based WCPM scores as follows.

In Step 1, the passage parameters are estimated and equated (i.e., calibrated on the same scale) as demonstrated in Eqs. 1 and 2. In the current study, we performed a concurrent calibration under an NEG common-passage design, where all passage parameters from the combined data were concurrently estimated by treating missing observations from unassigned passages as missing data. As a result, this step would establish a pool of passages with passage parameters calibrated on the same scale.

In Step 2, data from students are collected by using selected passages from Step 1. Then, the accuracy parameter θ and the speed parameter τ are estimated as shown in Eqs. 1 and 2, assuming the passage parameters are known based on the performed calibration in Step 1. Also, variance of τ and the covariance between τ and θ that estimated from Step 1 are treated as known parameter values in this step. Since the model parameters for passages are already equated in

Step 1, the estimated θ and τ are comparable between students, as well as within students across multiple testing occasions, regardless of what set of passages they read from the equated passage pool.

In Step 3, a set of passages is selected from the calibrated passage pool (created in Step 1) for the purpose of computing the model-based WCPM scores. We call this set of selected passages as the reference passages. The reference passages can be passages other than the student read to estimate θ and τ, as long as the reference passages and the passages the student read are from the same calibrated passage pool. Note that this step is necessary to obtain WCPM scores in a common scale, unless all students have read the same set of passages. Since the scale of WCPM scores is passage dependent, model-based WCPM scores would not be equated without a reference passage set, although the estimated θ and τ in Step 2 are equated.

In Step 4, equated model-based WCPM scores are derived from the estimated θ, τ, and the passage parameters of the reference passages (see *f* in Eq. 3). Equated model-based WCPM scores would be obtained as a result of equated θ and τ in Step 2 and the use of a reference passage set in Step 3. Note that a model-based WCPM score depends on $E[U_{\cdot j}]$ and $E[T_{\cdot j}]$ (Eqs. 4 and 5), which further depend on what passages are selected as the reference passages.

Equating Observed WCPM Scores with Traditional Methods

Equating observed WCPM scores from different passages has been mostly done by the mean, linear, and equipercentile equating methods (e.g., Albano & Rodriguez, 2012; Santi et al., 2016; Stoolmiller et al., 2013). This study focused on linear and equipercentile equating methods to be compared with the model-based approach.

When applying the observed score equating methods, appropriate equating designs are necessary to control the confounding effects from differences between groups who take two test

forms, *X* and *Y*, on the difficulty estimates of the two test forms. The equivalent group (EG) design assumes the groups are sampled from the same target population *T*. Thus, the groups are considered to be randomly equivalent. This design assumes no confounding effects from non-equivalency of the two groups. On the other hand, the NEG design assumes the two groups are sampled from two different populations, *P* and *Q*. Therefore, the confounding effects due to differences between the groups should be controlled statistically, which is usually achieved by the nonequivalent group with anchor test (NEAT) design. Specifically, an anchor test *V* is taken by both groups and the scores from *V* are used to control the differences between groups.

In the context of ORF assessments, the "groups" refer to samples of students that read different sets of passages, which are analogous to "test forms". Also, the "anchor test" is analogous to a set of common passages that are read by both groups of students. Below, we provide brief descriptions of the linear and equipercentile equating methods. Readers are referred to Santi et al. (2016) for a more detailed information about these methods as well as their application to observed WCPM score equating.

Linear Equating. Linear equating is essentially a linear conversion that sets the standardized deviation scores to be equal for the two forms. Let $\mu(X)$ and $\mu(Y)$ be the score means and $\sigma(X)$ and $\sigma(Y)$ be the standard deviation scores for Form *X* and Form *Y*, respectively. For linear equating with the EG design, the formula for the linear conversion is

$$l_Y(x) = \sigma(Y)\left[\frac{x - \mu(X)}{\sigma(X)}\right] + \mu(Y). \quad (6)$$

For linear equating with the NEG design, the formula for the linear conversion becomes

$$l_{Y_s}(x) = \frac{\sigma_s(Y)}{\sigma_s(X)}[x - \mu_s(X)] + \mu_s(Y), \quad (7)$$

where *s* indicates the synthetic population.

Equipercentile Equating. Equipercentile equating maps scores on Form *X* that have the same percentile ranks to scores on Form *Y*. For the EG design, Braun and Holland (1982) indicated that the equipercentile equating function is $\boldsymbol{e_Y(x) = G^{-1}[F(x)]}$, where $\boldsymbol{G^{-1}}$ is the inverse of the cumulative distribution function *G*. For the NEG design, the equipercentile function for the synthetic population is $\boldsymbol{e_{Y_S}(x) = Q_S^{-1}[P_S(x)]}$, where $\boldsymbol{P_S}$ is the percentile rank function for Form *X*, while $\boldsymbol{Q_S^{-1}}$ is the percentile function.

Simulation Study

A simulation study was conducted to evaluate the performance of the model-based equating method in comparison to the observed-score linear and equipercentile equating methods for equating WCPM scores.

Equating Data Collection Design

The current study assumed the NEAT design (Kolen & Brennan, 2004) for observed and model-based score equating. Each group was assumed to read a unique set of passages and a set of common passages, which were used to estimate passage parameters by linking two groups. Common passages were treated as external anchors. In other words, they were not part of the WCPM score estimations. In addition, observed-score equating methods that assumed the EG condition were included to evaluate the impact of the violation of the group equivalence assumption.

Under the NEAT design, Tucker's linear method and the frequency estimation equipercentile method were used to estimate the synthetic population parameters in the equating functions. The two methods differ in terms of their statistical assumptions and complexity. Tucker's linear method is simpler and is expected to perform better with small sample sizes,

while frequency estimation equipercentile requires large sample sizes to yield accurate parameter estimates.

Simulation Conditions

We assumed two groups of students, where one group read a set of 6 easy passages (referred to as Group E, hereafter) and the other group read a set of 6 hard passages (referred to as Group H, hereafter) based on the time intensity and accuracy difficulty parameters. In addition to group-specific passage sets, Group E and Group H were also assumed to have read a set of common passage(s) with medium time intensity and difficulty levels. Common passage(s) were treated as external anchors, namely, they were not part of the ORF scoring. Three manipulated factors in the simulation study were (a) the sample size per group (100, 300, or 500), (b) the number of anchor passages (1 or 3), and (c) population ORF ability discrepancy controlled by speed and accuracy parameters (no discrepancy, small or large). Note that the condition with no population ORF ability discrepancy is the condition where the EG assumption holds (i.e., random equivalence). By crossing all three factors, 18 total conditions were identified for data generations.

These simulation factors were selected due to their relevance to anticipated equating methods performance (model-based and/or observed-score methods). We also aimed to identify realistic conditions in order to draw more generalizable conclusions. Specifically, the chosen levels of the sample size reflect small to large sample sizes in ORF assessments reported in empirical studies (e.g., Nese et al., 2015). In addition, equipercentile equating is known to require larger samples for more accurate ORF score equating (e.g., Santi et al., 2016). The number of anchor passages is expected to affect the performance of the observed-score and model-based equating methods with the NEAT design. Population discrepancy is expected to

affect the estimation accuracy of the observed-score equating methods with the EG design. It may also affect the performance of the model-based and observed-score equating methods with the NEAT design.

Data Generation

True passage parameters were selected from a calibrated passage pool as part of a previously conducted study (Potgieter et al., 2017). ORF data were collected for 150 passages, where approximately 150 students read each passage on average. The selection of the easy and hard passage sets was performed by the inspection of time intensity and difficulty parameters of the calibrated passages. After excluding some outlier values, easy passages were identified from the calibrated passage pool which had difficulty and time intensity parameters close to the lowest values. Similarly, hard passages were ones which had calibrated parameter values close to the highest values in the passage pool. Common passages were also selected from the pool with medium-level difficulty and intensity parameters. Also, during the passage selection, we prioritized passages with close to 50 words to control a possible impact of different passage lengths. The average difference between easy and hard passage sets were .651 and .180 for difficulty and time intensity parameters, respectively. Actual passage parameters and the number of the words for all passages are provided in Table 1.

Speed and accuracy ability parameters (θ and τ) were generated from the multivariate normal distribution by altering means, depending on the population discrepancy for each condition. On the other hand, variances of speed ability and the covariance between speed and accuracy ability were fixed by using the values obtained from the formerly calibrated passage pool. Values of the model hyperparameters in different simulation conditions are summarized in Table 2. Fifty data sets were generated for each of the 18 simulation conditions. Reading time in

seconds and the numbers of words read correctly per passage were generated by the latent-variable ORF model described earlier (Kara et al., 2020).

Equating Procedures and Analyses

The WCPM scores from Group E are equated to those from Group H. Observed WCPM scores were computed using the generated reading time and correctly read word count data, by dividing the total number of words read correctly by the total reading time in seconds and multiplied by 60. The WCPM scores were rounded to integers before applying the observed-score equating procedures to be consistent with WCPM score reporting in practice. Under the NEG assumption, Tucker's linear and frequency estimation equipercentile equating methods with loglinear presmoothing (Holland & Thayer, 2000) were performed. Under the EG assumption, linear and presmoothed equipercentile equating methods were performed.

For the model-based equating, we followed the steps described earlier in the previous section. Note that a concurrent calibration of passage parameters was performed for each replication of the simulation. Thus, equating errors associated with passage parameters were incorporated into the process. Then, model-based WCPM scores were obtained using Eq. 3, where person parameters (i.e., accuracy and speed) were estimated using the same Bayesian MCMC technique as in Kara et al. (2020). More specifically, we first estimated accuracy and speed parameters (θ and τ) of simulated students in Group E for each generated dataset (for easy passages). Then, we used these estimated θ and τ to obtain the model-based WCPM scores with passage parameters in the hard passage set. As a result, the model-based WCPM scores were equated to the scale of scores in Group H, because all WCPM scores were computed for the same set of hard passages.

As per the Bayesian estimation, we set the number of iterations as 40000, the number of burning as 20000, and number of thinning as two to alleviate autocorrelations. Three chains were run with initial values set as $\sigma_{\theta\tau}$ = -0.5, $\sigma_{\tau}^{-2}|\theta_j$= 30; $\sigma_{\theta\tau}$ = -0.1, $\sigma_{\tau}^{-2}|\theta_j$ = 60; $\sigma_{\theta\tau}$ = 0.1, $\sigma_{\tau}^{-2}|\theta_j$ = 20. Model convergence was assessed by the potential scale reduction factor (PSRF; Brooks & Gelman, 1998) with PSRF < 1.1 indicating adequate convergence. In addition, the effective sample size (ESS) > 400 was used as a rule to indicate satisfactory precision of the Bayesian estimation (Zitzmann & Hecht, 2019).

Observed-score linear and equipercentile equating were performed by using the R package *equate* (Version 2.0-5; Albano, 2016). As per model-based equating, we used R package *R2jags* (Su & Yajima, 2015) to interface with JAGS (version 4.3.0; Plummer, 2015). JAGS syntax for the passage calibration and final equating step with fixed parameter values are provided in the Appendix. We used R (Version 4.1.2; R Core Team, 2016) for all our analyses and visualizations.

In summary, we performed four observed-score equating methods, namely, linear and equipercentile equating with EG and NEG (i.e., NEAT) designs. As mentioned earlier, methods with the EG assumption were included in the simulation to demonstrate the effects of their misuse, when the group equivalency assumption is not met. The performance of these observed-score equating methods were compared to the model-based equating under the 18 conditions elaborated above.

Evaluation Criteria

As indicated earlier, the direction of equating was from Group E to Group H. In other words, WCPM scores from easy passages were equated to hard passages, as if Group E read the hard passages. Thus, we examined the degree to which equated ORF scores of Group E was

close to their true (i.e., population) values. Population WCPM scores of Group E on the hard passages were computed by using Eq. 3 with the population speed and accuracy abilities (θ and τ) and the true values of the hard passages' parameters. Thus, equated WCPM scores were compared to these population (i.e., true) WCPM values for the hard passages.

Equating errors were evaluated using the following three measures: absolute relative bias (*ARB*), standard error (*SE*), and root mean squared error (*RMSE*). Given *R* replications and true WCPM score f_j from person *j*, the *ARB* of WCPM score estimate $\hat{f}$ was calculated as:

$$ARB(f_j) = \left| \frac{\frac{1}{R}\sum_{r=1}^{R} \hat{f}_j - f_j}{f_j} \right|. \quad (8)$$

The *SE* of WCPM score estimate $\hat{f}$ was calculated as:

$$SE(f_j) = \sqrt{\frac{1}{R}\sum_{r=1}^{R}\left(\hat{f}_j - \frac{\sum_{r=1}^{R}\hat{f}_j}{R}\right)^2}. \quad (9)$$

Finally, the *RMSE* of WCPM score estimate $\hat{f}$ was calculated as:

$$RMSE(f_j) = \sqrt{\frac{1}{R}\sum_{r=1}^{R}(\hat{f}_j - f_j)^2}. \quad (10)$$

The averages of the above outcome measures across persons were computed for the group which took the easy passages (Group E) to evaluate the impacts of the manipulated factors.

Results

Results from the simulation study are summarized in terms of the average *ARB*, *SE*, and *RMSE* values for the equated WCPM scores for Group E (the group who took easy passages) in Figures 1 - 3. Results for the observed-score equating methods with EG design were not shown

in the figures due to large recovery index values compared to methods with NEAT design. Rather, we provided our interpretations of the results for observed-score equating methods with EG design in relevant paragraphs. In all conditions, the ORF models' estimation converged with PSRF < 1.1 and ESS > 400 for all person parameters in all replications.

The average *ARB* values for the equated WCPM scores with the three equating methods that assume the NEAT design are graphically summarized in Figure 1. As expected, the large population discrepancy led to higher average *ARB* for all three methods. Under all simulation conditions, model-based equating outperformed the linear and equipercentile equating methods. The difference in average ARB values was larger between the observed and model-based methods under the large population discrepancy condition. Having more anchor passages contributed to lower the average bias for all three methods, especially when the difference between groups' ORF ability was large. The effect of sample size, however, was not as prominent as the other simulation factors. In other words, we did not observe a consistent decrease in average ARB values by increasing the sample size. On the other hand, the effect of wrongly assuming group equivalence with linear and equipercentile equating methods was severe. Although not plotted in Figure 1, the average ARB values for these methods that assumed equivalent groups were much higher than the scale of the vertical axis in Figure 1 can capture: over 0.2 and 0.4 with small and large levels of population discrepancy, respectively.

Note that the ARB values evaluated above were the averages across all observations in each simulation condition. These averages were below .05 for all three methods with no and small population discrepancy conditions. Also, they were all below .05 for all three methods under large population discrepancy conditions with three anchor passages. With a large population discrepancy and a single anchor passage, only model-based method displayed the

average ARB below .05 with all sample sizes. To further examine the performance of the three equating methods that assumed the NEAT design, we provided the visuals for individual ARB values per sample size in the supplementary material.

The average *SE* values of the equated WCPM scores are summarized in Figure 2. As expected, larger sample sizes led to smaller average *SE* for the observed-score equating methods with the NEAT design. Furthermore, a large population discrepancy led to higher average *SE* for all three equating methods, yet the effect of the number of anchor passages was not prominent. Average SE values were always lower for the model-based approach under all simulation conditions. Although not shown in Figure 2, for the two observed-score equating methods with the EG assumption, the average SE values were close to their counterparts that assumed the NEAT design, when there was no population discrepancy. Interestingly, under the conditions with a large population discrepancy, the two observed-score methods with the EG assumption had lower average *SE* values (around 5.5 for all sample sizes and number of anchor passages), which may be counterintuitive. Our interpretation was that it was because they performed consistently worse, due to an incorrect assumption about group equivalency.

The average *RMSE* values of the equated WCPM scores are presented in Figure 3. Overall, the pattern was similar to the results for average *ARB*, which is an indication that a majority of errors were due to systematic errors (bias), rather than sampling fluctuations (*SE*), except that *RMSE* displayed some impacts of sample sizes attributed to the *SE*s, especially for the equipercentile equating. Larger sample sizes would be needed for the equipercentile equating to perform as good as the linear equating, when there was a non-zero population discrepancy. The effect of number of anchor passages was not prominent yet a slight decrease was observed with the increase of number of anchor passages, when the population discrepancy was large. In

sum, the model-based equating method performed consistently better than other equating methods in all simulation conditions in terms of average *RMSE*. Lastly, the average RMSE for the observed-score methods with EG assumption were much larger than the scale of the vertical axis in Figure 3: 15 and 35 for the conditions with small and large population discrepancy, respectively. Thus, the wrong assumption of the group equivalency by the observed-score equating methods resulted in substantially larger total equating errors, compared to methods that assumed the NEAT design.

Discussion

In this paper, we demonstrated the model-based approach for equating WCPM scores in ORF assessments based on a latent-variable measurement model (Kara et al., 2020). We conducted a simulation study to evaluate the performance of the model-based equating method in comparison with traditionally used observed-score equating methods, linear and equipercentile equating.

Overall, the results demonstrated that the model-based approach performed satisfactorily well under all simulation conditions. This was encouraging for the use of the model-based approach to equate WCPM scores. For example, Babcock and Hodge (2020) showed the utility of the Rasch model for performing equating on traditionally-scored exams with relatively low sample sizes. Their findings are in line with ours: we demonstrated that the latent-variable model-based approach performed as good as or better than the traditional observed-score equating methods. On the other hand, the performance of the observed-score equating methods depended on specific conditions. This was not surprising since it is known that the accuracy of the traditional equating methods depends on the degree to which the underlying assumptions, such as linearity for the linear equating, are met (Albano & Rodriguez, 2012). Nevertheless, the

observed-score equating methods performed equally well as the model-based approach, when the discrepancy was small or non-existent in terms of the population ORF abilities between groups.

In addition, the results demonstrated dramatic impacts of the EG assumption for the observed-score equating methods. These methods did not perform well, unless the EG assumption was met in the data. Since the assumption of having equivalent groups would be realistically hard to ensure in real-life conditions, this strong assumption required by the classical equating methods are likely to be violated. This was also pointed out by Albano and Rodriguez (2012), stating that the traditional WCPM scores as part of classroom-based measurements were not designed to meet such assumptions. Therefore, it is paramount to employ an equating method that incorporates the assumption of group nonequivalence, such as an observed-score equating method with a NEAT design or the model-based equating approach, which does not rely on such an assumption, as demonstrated in this paper. In addition to a horizontal equating, these methods should be preferred especially for vertical equating, where groups are comprised of students from different grade-levels, where population discrepancies in ORF speed and accuracy abilities naturally exist. Thus, adopting an equating method with a NEAT design or the model-based approach would be an optimal choice for WCPM score equating for the purpose of progress monitoring over multiple years.

Equating ORF assessment scores is essential to ensure score comparability both between and within students. Besides its better overall performance in terms of more accurate equated scores, the model-based equating approach for WCPM scores potentially provides several other practical advantages to researchers and practitioners. First, a calibrated passage pool allows practitioners to build reading passage forms of various difficulties. It is also important to note that building such a calibrated passage pool would not require a complete design where all

students are expected to read all passages. This is thought to be an important aspect of adopting a model-based approach not only for equating studies but also for the measurement of ORF ability. Second, similar to the advantage of IRT-based test scoring, having a calibrated passage pool can be a basis for a potential future development of pre-equated ORF assessment forms and a computer adaptive test version of an ORF assessment.

In addition to its advantages in scoring and scaling, it has been demonstrated that the model-based approach to ORF assessment data has an advantage over the traditional observed-score approach, because it allows the computation of standard errors for each estimated WCPM score, namely, conditional standard errors of measurement (*CSEM*; Nese & Kamata, 2021). On the other hand, with the observed WCPM scores, only one equivalent quantity can be computed for the entire sample of students (namely, *SEM*: standard error of measurement). Based on these additional advantages of the model-based approach to ORF assessment data and the results from our study, it is recommended that model-based equating method to be considered for equating WCPM scores.

There are several limitations in our study, mainly associated with the simulation we conducted. Specifically, the results from the simulation are limited to handful of factors manipulated that are sample size per group, number of anchor passages, and level of discrepancy in groups' ORF ability. The performance of the observed-score and model-based equating methods may differ with other levels of these factors and/or other factors that were not considered, such as the level of test difficulty discrepancy, which was a fixed factor. Nevertheless, it is worth noting that the two passage sets (easy and hard) were formed based on a previously calibrated passage pool. Thus, this level of difficulty discrepancy in the two sets of passages is expected to reflect a realistic condition. On the other hand, the number of unique

passages read by each group (i.e., the test length) was also a fixed factor. Moreover, we did not consider the varying lengths of passages by intentionally selecting passages with approximately 50 words. Future studies can focus on factors not considered here and relevant to other realistic ORF assessment conditions.

References

Albano, A. D. (2016). *equate: Observed-score linking and equating* (Version 2.0-7) [Computer software]. https://CRAN.R-project.org/package=equate.

Albano, A. D., & Rodriguez, M. C. (2012). Statistical equating with measures of oral reading fluency. *Journal of School Psychology*, *50*(1), 43-59. https://doi.org/10.1016/j.jsp.2011.07.002

Albano, A. D., Christ, T. J., & Cai, L. (2018). Evaluating equating in progress monitoring measures using multilevel modeling. *Measurement*, *16*(3), 168-180. https://doi.org/10.1080/15366367.2018.1483663

Babcock, B., & Hodge, K. J. (2020). Rasch versus classical equating in the context of small sample sizes. *Educational and Psychological Measurement*, *80*(3), 499-521. https://doi.org/10.1177/0013164419878483

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

Braun, H. I., & Holland, P. W. (1982). Observed-score test equating: A mathematical analysis of some ETS equating procedures. In P.W. Holland & D. B. Rubin (Eds.), *Test equating* (pp. 9-49). New York: Academic.

Brooks, S. P., & Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, *7*, 434-455. https://doi.org/10.1080/10618600.1998.10474787

Chaparro, E. A., Stoolmiller, M., Park, Y., Baker, S. K., Basaraba, D., Fien, H., & Smith, J. L. M. (2018). Evaluating passage and order effects of oral reading fluency passages in second grade: A partial replication. *Assessment for Effective Intervention*, *44*(1), 3-16. https://doi.org/10.1177/1534508417741128

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. https://doi.org/10.1080/02796015.2007.12087956

Cummings, K. D., Park, Y., & Schaper, H. A. B. (2013). Form effects on DIBELS next oral reading fluency progress- monitoring passages. *Assessment for Effective Intervention*, *38*(2), 91-104. https://doi.org/10.1177/1534508412447010

Francis, D. J., Santi, K. L., Barr, C., Fletcher, J. M., Varisco, A., & Foorman, 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., 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

Good, R. H., & Kaminski, R. A. (2002). *DIBELS oral reading fluency passages for first through third grades (technical report no. 10)*. https://dibels.uoregon.edu/sites/dibels1.uoregon.edu/files/DORF_Readability.pdf

Hasbrouck, J., & Tindal, G. (2006). Oral reading fluency norms: A valuable assessment tool for reading teachers. *The Reading Teacher*, *59*(7), 636-644. http://www.jstor.org/stable/20204400

Holland, P. W., & Thayer, D. T. (2000). Univariate and bivariate loglinear models for discrete test score distributions. *Journal of Educational and Behavioral Statistics*, *25*, 133-183. https://doi.org/10.2307/1165330

Kara, Y., Kamata, A., Potgieter, C., & Nese, J. F. (2020). Estimating model-based oral reading fluency: A Bayesian approach. *Educational and Psychological Measurement, 80*(5), 847-869. https://doi.org/10.1177/0013164419900208

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking*. New York, NY: Springer.

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. (2014-2018). *Measuring Oral Reading Fluency: Computerized Oral Reading Evaluation* (Project No. R305A140203) [Grant]. Institute of Education Sciences, U.S. Department of Education. https://ies.ed.gov/funding/grantsearch/details.asp?ID=1492

Nese, J. F. T., Kamata, A., & Alonzo, J. (2015, July). *Exploring the evidence of speech recognition and shorter passage length in computerized oral reading fluency* [Conference presentation]. Meeting of the Society for the Scientific Study of Reading, Kailua-Kona, HI, United States.

Nese, J. F. T., & Kamata, A. (2021). Addressing the large standard error of traditional CBM-r: Estimating the conditional standard error of a model-based estimate of CBM-r.

Assessment for Effective Intervention, *47*(1), 53-58. https://doi.org/10.1177/1534508420937801

Plummer, M. (2015). JAGS: *A program for analysis of Bayesian graphical models using Gibbs sampling* (Version 4.0.0) [Computer software]. https://sourceforge.net/projects/mcmc-jags/

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

R Core Team. (2016). *R: A language and environment for statistical computing* (Version 4.1.2) [Computer software]. Vienna, Austria: R Foundation for Statistical Computing. http://www.R-project.org/

Santi, K. L., Barr, C., Khalaf, S., & Francis, D. J. (2016). Different approaches to equating oral reading fluency passages. In K. D. Cummings & Y. Petscher (Eds.), *The fluency construct: Curriculum-based measurement concepts and applications* (pp. 223-265). Springer Science + Business Media. https://doi.org/10.1007/978-1-4939-2803-3_9

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*, 76-90. http://dx.doi.org/10.1177/1534508412456729.

Su, Y. S., & Yajima, M. (2015). *R2jags: Using R to run 'JAGS'* (Version 0.5–7) [Computer software]. https://cran.r-project.org/web/packages/R2jags/index.html

van der Linden, W. J. (2007). A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika*, *72*, 287-308. https://doi.org/10.1007/s11336-006-1478-z

White, S., Sabatini, J., Park, B. J., Chen, J., Bernstein, J., and Li, M. (2021). *The 2018 NAEP Oral Reading Fluency Study (NCES 2021-025)*. U.S. Department of Education.

Washington, DC: Institute of Education Sciences, National Center for Education Statistics. https://nces.ed.gov/nationsreportcard/subject/studies/orf/2021025_orf_study.pdf

Yi, E. (2021). Impact of Passage Effects on Oral Reading Fluency Administration and Scoring (Publication No. 28412749) [Doctoral dissertation, Northern Illinois University]. ProQuest Dissertations & Theses Global.

Zitzmann, S., & Hecht, M. (2019). Going beyond convergence in Bayesian estimation: Why precision matters too and how to assess it. *Structural Equation Modeling: A Multidisciplinary Journal*, *26*, 646-661. https://doi.org/10.1080/10705511.2018.1545232

Tables

Table 1

True Passage Parameter Values

Passage Difficulty	a	b	α	β_0	Number of Words
Easy	0.500	-2.771	6.176	1.723	47
	0.444	-2.923	5.747	1.768	49
	0.557	-2.760	4.841	1.711	50
	0.505	-2.856	5.064	1.750	50
	0.447	-2.949	4.408	1.705	50
	0.476	-2.877	3.936	1.752	49
Hard	0.569	-2.270	5.808	1.963	47
	0.569	-2.370	4.559	1.972	54
	0.603	-2.072	5.566	1.879	50
	0.618	-2.164	4.239	1.863	49
	0.571	-2.304	3.795	1.894	49
	0.588	-2.049	5.248	1.916	50
Medium*	0.575	-2.425	6.821	1.806	54
	0.573	-2.479	4.726	1.810	49
	0.544	-2.423	4.648	1.812	50

**Common passages. First set of parameters were used for conditions with one anchor passage.*

Table 2

True Person Hyperparameter Values

		Group E	Group H
Population Discrepancy	No (Eq. Groups)	$\begin{pmatrix}\theta\\ \tau\end{pmatrix} \sim MVN \begin{pmatrix}0\\ 0\end{pmatrix} \begin{pmatrix}1 & .16\\ .16 & .16\end{pmatrix}$	$\begin{pmatrix}\theta\\ \tau\end{pmatrix} \sim MVN \begin{pmatrix}0\\ 0\end{pmatrix} \begin{pmatrix}1 & .16\\ .16 & .16\end{pmatrix}$
	Small	$\begin{pmatrix}\theta\\ \tau\end{pmatrix} \sim MVN \begin{pmatrix}.25\\ .1\end{pmatrix} \begin{pmatrix}1 & .16\\ .16 & .16\end{pmatrix}$	$\begin{pmatrix}\theta\\ \tau\end{pmatrix} \sim MVN \begin{pmatrix}-.25\\ -.1\end{pmatrix} \begin{pmatrix}1 & .16\\ .16 & .16\end{pmatrix}$
	Large	$\begin{pmatrix}\theta\\ \tau\end{pmatrix} \sim MVN \begin{pmatrix}.5\\ .2\end{pmatrix} \begin{pmatrix}1 & .16\\ .16 & .16\end{pmatrix}$	$\begin{pmatrix}\theta\\ \tau\end{pmatrix} \sim MVN \begin{pmatrix}-.5\\ -.2\end{pmatrix} \begin{pmatrix}1 & .16\\ .16 & .16\end{pmatrix}$

Note. Group E and Group H are the two groups that are assumed to read a set of easy or hard passages, respectively.

Figures

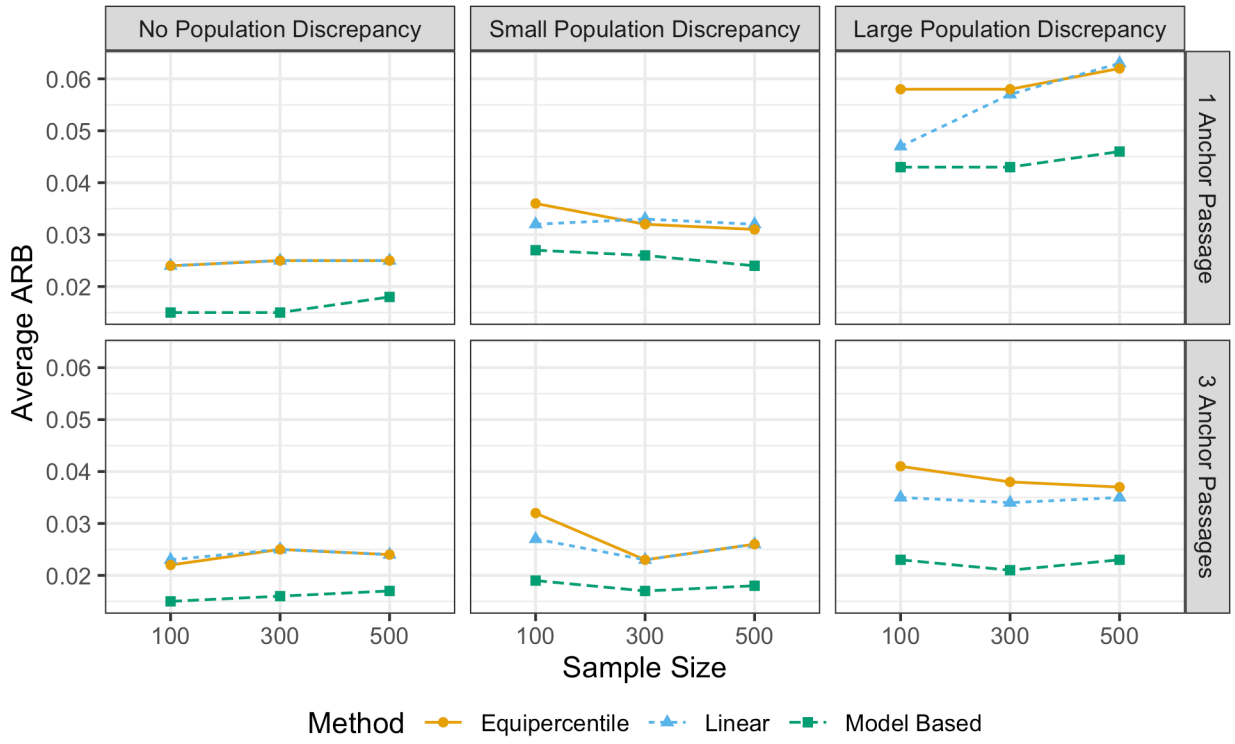

Figure 1. Average absolute relative bias (ARB) of the equated WCPM scores

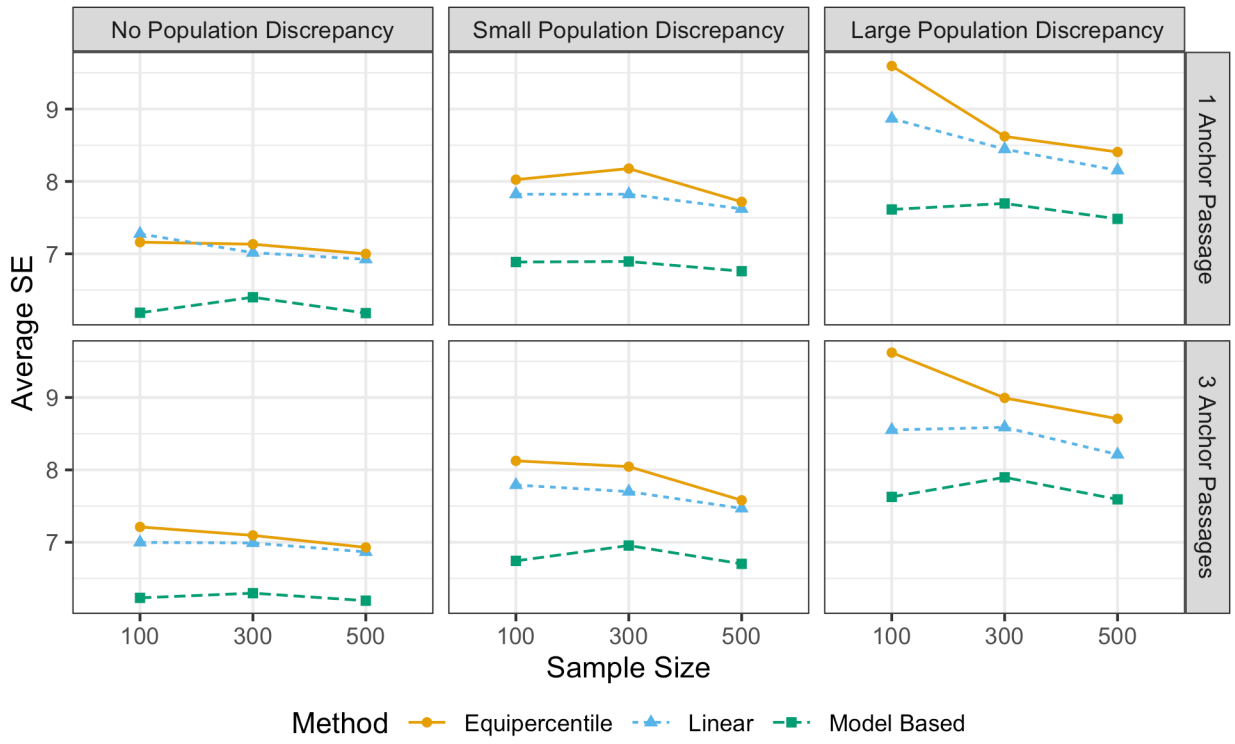

Figure 2. Average standard error (SE) of the equated WCPM scores

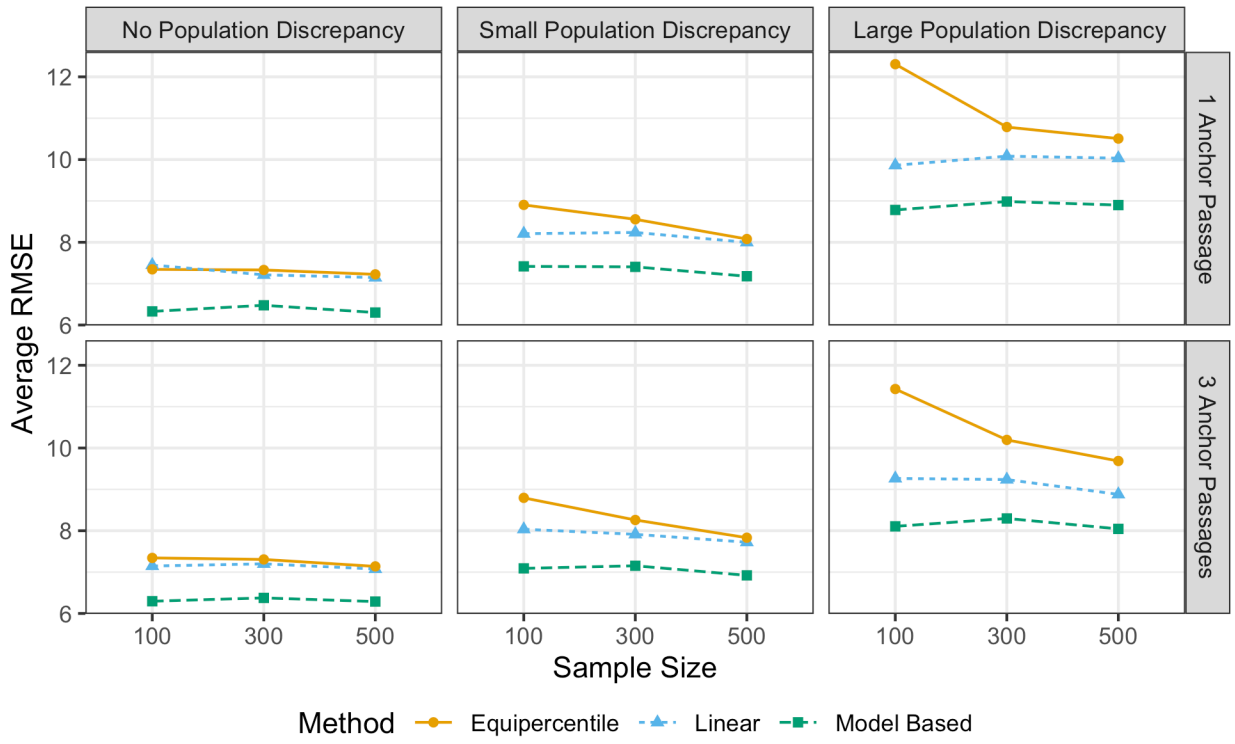

Figure 3. Average root mean square error (RMSE) of the equated WCPM scores

Appendix

JAGS Syntax for the Concurrent Calibration Step

```
model{
    for (j in 1:J){
    for (i in 1:I){
    res[j,i]~dbin(p[j,i], nw[j,i])
    logit(p[j,i]) <- a[i]*(theta[j]-b[i])
    log_tim[j,i]~dnorm(mu[j,i], prec.t[i]) #variance=1/alpha^2, so precision=alpha^2
    mu[j,i] <- beta[i]-tau[j]
    }
    theta[j]~dnorm(0,1)
    tau[j]~dnorm(mtau[j], ptau) #distribution of tau conditional on theta
    mtau[j] <- cvr*theta[j] #cvr is the covariance between tau and theta
    }
    for(i in 1:I){
    prec.t[i] <- pow(alpha[i], 2)
    alpha[i] ~dnorm(0, 0.01) T(0,)
    b[i]~dnorm(0, 0.01)
    beta[i]~dnorm(0, 0.01)
    a[i]~dnorm(0, 0.01) T(0,)
    }
    ptau~dgamma(0.01, 0.01) #conditional precision of tau
    vtau <- 1/ptau
    tau.var <- vtau + (pow(cvr,2)) #tau.var is the variance of tau's marginal distribution
    cvr~dnorm(0, 0.01)
    crl <- cvr/sqrt(tau.var)
    }
```

JAGS Syntax for the Final Step of the Model-Based Equating

Note that in the syntax below, all passage parameters, the variance of tau, and the covariance between theta and tau are treated as known quantities and supplied along with data values during the JAGS estimation.

```
model{
    for (j in 1:J){
    for (i in 1:I){
    res[j,i]~dbin(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]
    }
```

```
theta[j]~dnorm(0, 1)
tau[j]~dnorm(mtau[j], ptau)
mtau[j] <- cvr*theta[j]
}
for(i in 1:I){
 prec.t[i] <- pow(alpha[i],2)
}
}
```

Supplementary Material

Absolute relative bias (ARB) values for individual true word read correctly per minute (WCPM) scores are plotted in Figures S1-S3. Note that the horizontal line reflects the .05 threshold value for reference, which is considered to be the acceptable threshold in simulation studies (Hoogland & Boomsma, 1998). All values below this threshold were faded out for ease of interpretation. It was observed that the equipercentile equating had larger ARB values for the extreme low and high WCPM scores compared to linear and model-based methods with the nonequivalent group with anchor test (NEAT) design. It is also worth noting that ARB yielded by the linear and equipercentile equating methods was larger for true scores that were at the tails of the score distribution. On the other hand, the performance of the model-based method was less affected by specific true score levels.

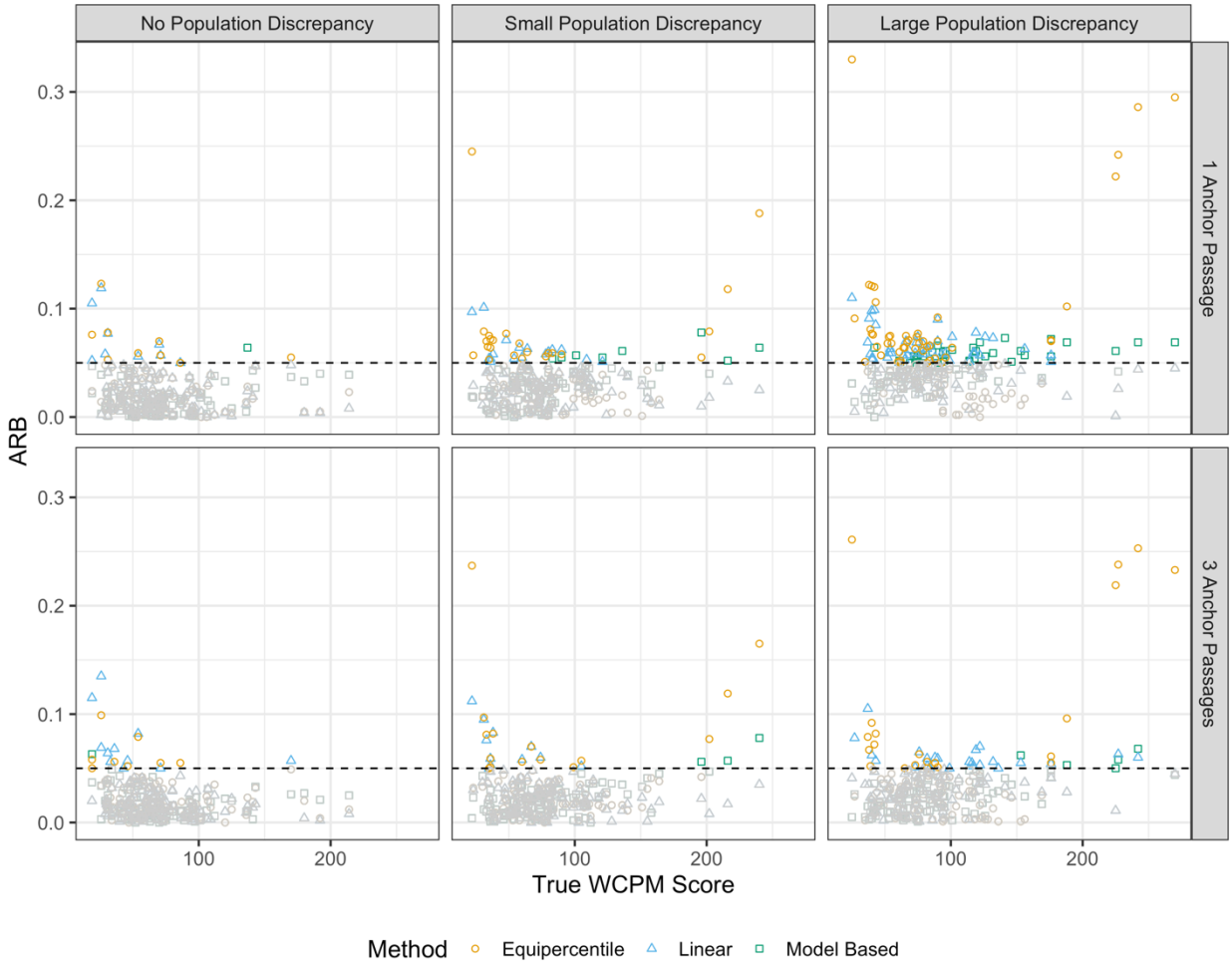

Figure S1. Absolute relative bias (ARB) for equated WCPM scores under conditions with N = 100

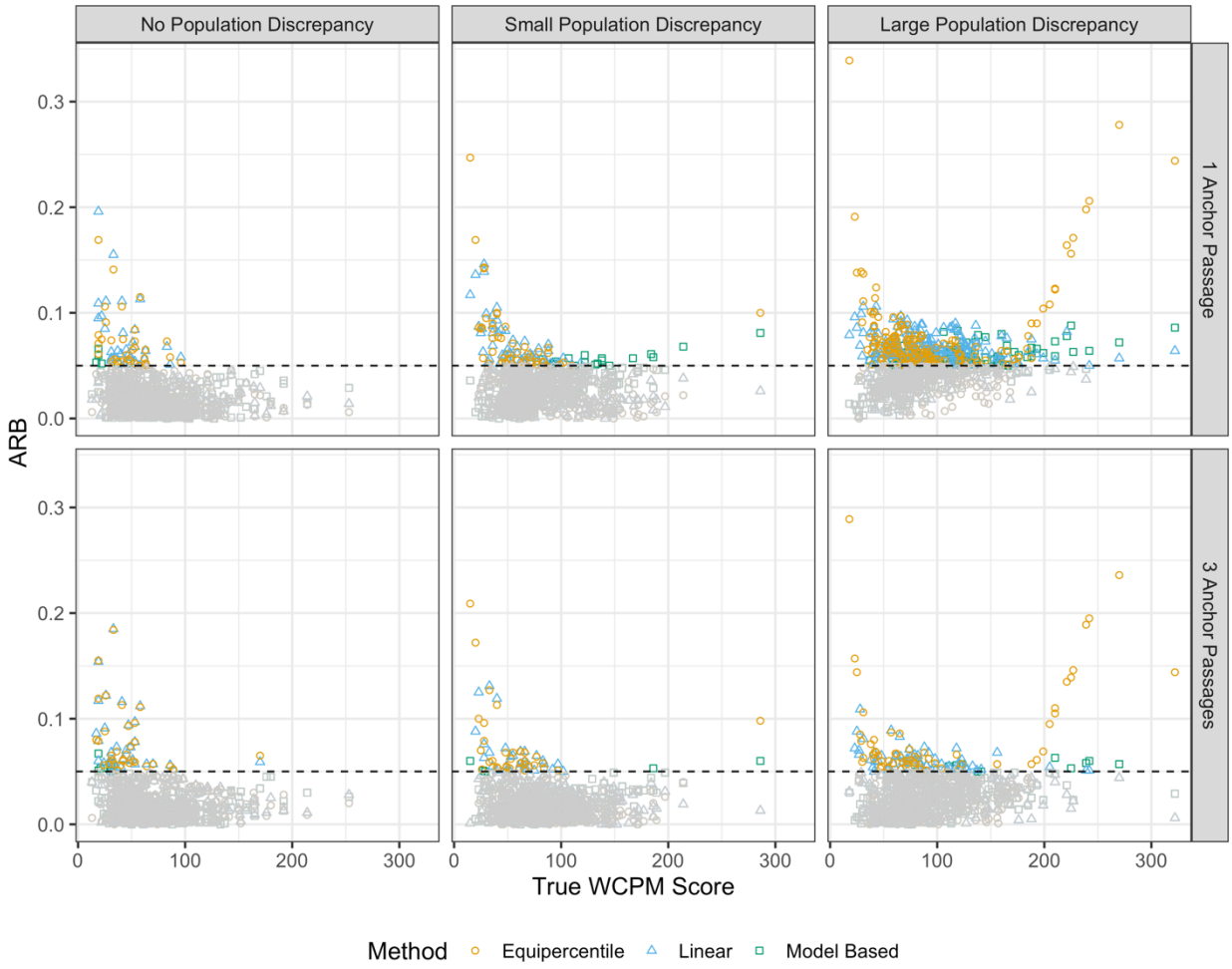

Figure S2. Absolute relative bias (ARB) for equated WCPM scores under conditions with N = 300

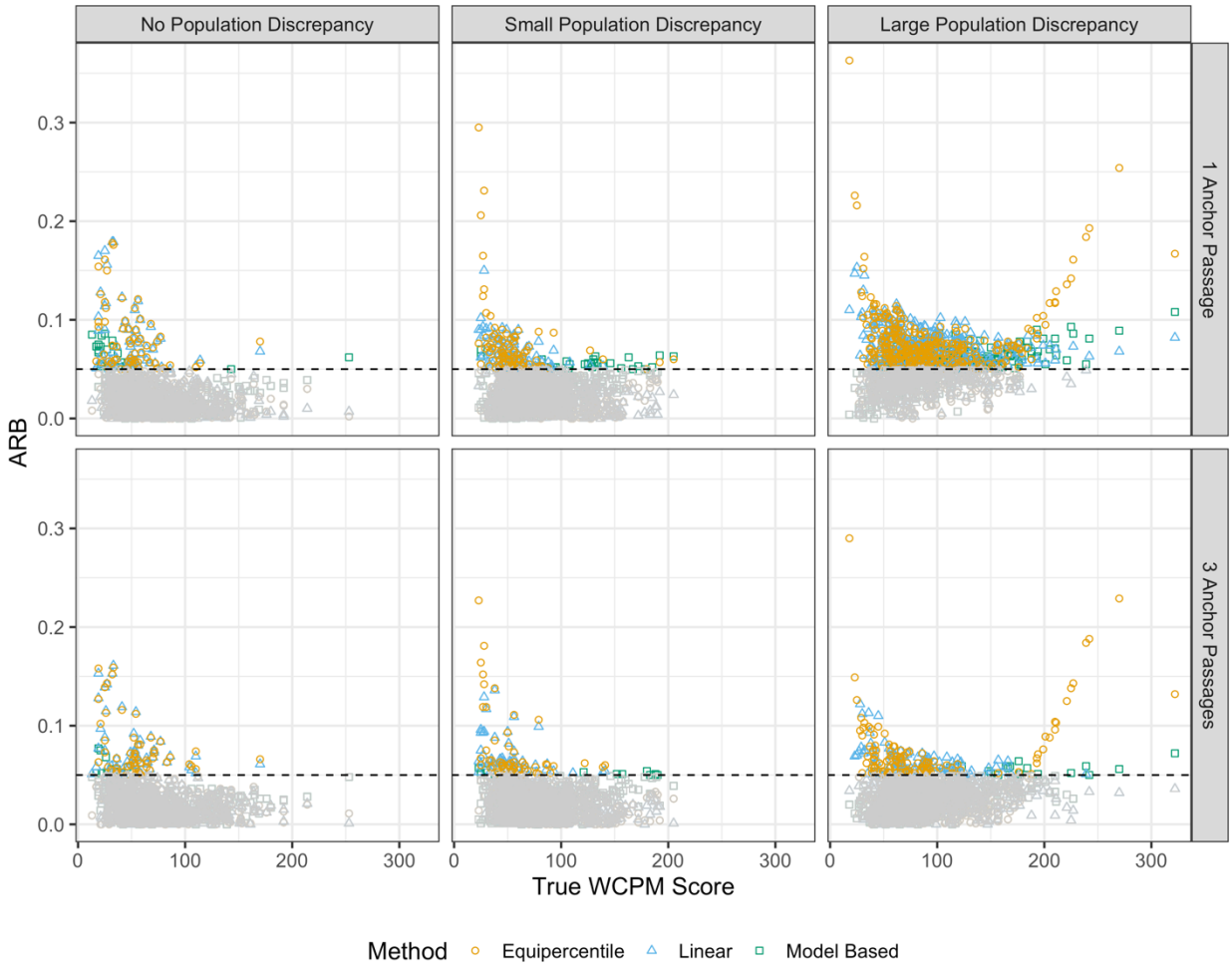

Figure S3. Absolute relative bias (ARB) for equated WCPM scores under conditions with N = 500

Hoogland, J. J., & Boomsma, A. (1998). Robustness studies in covariance structure modeling: An overview and a meta-analysis. *Sociological Methods & Research*, *26*(3), 329-367. https://doi.org/10.1177/0049124198026003003