Note: This is a conference paper presented at the International Meeting of Psychometrics Society (IMPS) annual meeting, July 2023, College Park, MD. This work is funded by Grant R305D200038 from Institute of Education Sciences - U.S. Department of Education.

# Asymptotic Standard Errors of Model-based Oral Reading Fluency Score Equating

Xin Qiao

University of South Florida

Akihito Kamata

Southern Methodist University

Cornelis Potgieter

Texas Christian University

Abstract

Oral reading fluency (ORF) assessments are commonly used to screen at-risk readers and to evaluate the effectiveness of interventions as curriculum-based measurements. As with other assessments, equating ORF scores becomes necessary when we want to compare ORF scores from different test forms. Recently, Kara et al. (2023) proposed a model-based equating method for ORF scores. However, they did not provide closed-form asymptotic standard errors (SEs) of the equated ORF scores while it is advocated to report SEs of equating in practice. Therefore, this study aims to address this remaining question. Specifically, the delta method was adopted to derive the asymptotic SEs of equated ORF scores. The ORF scoring was conducted using an approach that takes into account the calibration error. Its performance was compared to the standard practice that ignores calibration errors. It is expected that the two scoring approaches would further have impact on the recovery of equated ORF score SEs. A simulation study was conducted to evaluate the recovery of derived equating SEs in various conditions. Results suggested satisfactory recovery of the derived asymptotic SEs of equated ORF scores. In addition, taking into account the calibration error in ORF scoring can produce more accurate and precise equated ORF score SEs under larger sample size and longer test length.

# Asymptotic Standard Errors of Model-based Oral Reading Fluency Score Equating

## 1. Introduction

Oral reading fluency (ORF) assessments are commonly used as curriculum-based measurements to screen at-risk readers and to evaluate the effectiveness of interventions. Kara, Kamata, Potgieter, and Nese (2020) proposed the model-based word read correctly per minute (WCPM) scores as reliable score metric for oral reading fluency. Same as other assessments, equating ORF scores is necessary when comparing ORF scores from different test forms is of interest. More recently, Kara, Kamata, Qiao, Potgieter, and Nese (2023) proposed a model-based equating method for ORF scores. They demonstrated the benefits of the model-based equating method over traditional observed-score approaches. However, they did not provide closed-form asymptotic standard errors ($SE$s) of the equated ORF scores. Therefore, we aim to address this remaining question in this article.

It is advocated to report $SE$s of equating as the standard practice in practical test equating (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, 2018). One common approach to obtain $SE$s for the IRT equating coefficients is the delta method (Ogasawara, 2000, 2001a, 2001b). Delta method is a general method for the calculation of $SE$s for transformed parameters that are functions of some other parameters with known asymptotic variances. It is based on Taylor series expansion and can achieve good approximation when the parameter estimates are close to their true values and the transformation function is differentiable at these true values with non-zero derivatives. In the current study, the model-based ORF score is a function of passage parameters and latent variables and delta method can be used to derive the asymptotic $SE$s of equated ORF scores.

Similar to IRT equating, model-based equating for ORF scores includes two crucial components: 1) placing passage parameters onto the same measurement scales through concurrent calibration; and 2) equating the ORF scores. Ignoring the random

errors in the calibration process, which is the standard practice in IRT scoring, can affect the estimation accuracy of IRT true scores (e.g., Mislevy, Wingersky, & Sheehan, 1994; Yang, Hansen, & Cai, 2012). In our context, this may further affect the accuracy of $SE$s of ORF score equating. Therefore, we are interested in investigating the impact of ignoring calibration error in ORF equating $SE$s.

The purpose of the current study is two-fold: 1) to derive the $SE$s of equated ORF scores using the delta method; 2) to compare the recovery of $SE$s of equated scores with or without considering calibration error in quantifying the uncertainty of latent scores and in the application of delta method. We follow the model-based equating procedure for ORF scores focusing on a nonequivalent group anchor test (NEAT) design proposed in Kara et al. (2023). Specifically, we assume there are two test forms, Form $V$ and Form $U$, which shared a set of external anchor items. Further, these two test forms are administered to two groups, Group 1 and Group 2, respectively. These two groups are nonequivalent in terms of latent variable distributions. That is, we assume the two groups come from different populations.

The remainder of the article is organized as follows. We first present the derivation of the asymptotic $SE$s of equated ORF scores. Then, we present the simulation study design and results. Lastly, we provide conclusions and recommendations on the usage of equated ORF scores in practice.

## 2. Asymptotic Standard Errors of Equated ORF Scores

In this section, we present the derivation of asymptotic $SE$s of equated ORF scores. We first review the model-based estimation of ORF scores introduced by Kara et al. (2020). Then, we describe the equating of ORF scores based on a NEAT design with a focus on model parameter estimation. Lastly, asymptotic $SE$s of the equated ORF scores are obtained based on the delta method.

### ORF Score Estimation

The estimation of ORF scores is based on a joint model consisting of an binomial model for number of words read correctly per passage and a lognormal model for

response times (RTs) used for reading each passage (Kara et al., 2020).

**Binomial Model for Count Data.** We assume reading each word by person $j$ is an independent trial within total number of trials being $N_i \in \mathbb{N}_0$, the total number of words in the passage $i$. Thus, the number of words read correctly in each passage $u_{ij}$ follows a binomial distribution:

$$P(u_{ij}|N_i, p_{ij}) = \binom{N_i}{u_{ij}} p_{ij}^{u_{ij}} (1 - p_{ij})^{N_i - u_{ij}}, \tag{1}$$

with success probability defined as the two-parameter normal-ogive model $p_{ij} = \Phi(a_i\theta_j - c_i)$, where $\theta_j \in \mathbb{R}$ is the latent ability of person $j$ and $a_i > 0$ and $c_i \in \mathbb{R}$ are discrimination and intercept parameters of passage $i$, respectively.

**Lognormal Model for Response Time Data.** We adopt the lognormal model (van der Linden, 2006) where RTs are assumed to have a lognormal distribution:

$$f(t_{ij}|\tau_j, \beta_i, \alpha_i) = \frac{\alpha_i}{t_{ij}\sqrt{2\pi}} \exp(-\frac{1}{2}[\alpha_i(\log t_{ij} - (\beta_i - \tau_j))]^2), \tag{2}$$

in which $t_{ij}$ denotes RT (in seconds) for passage $i$ and person $j$; $\tau_j \in \mathbb{R}$ is the latent speed for person $j$; $\beta_i \in \mathbb{R}$ is the time intensity for item $i$, indicating the labor required for that passage; and $\alpha_i > 0$ represents the time discrimination parameter. We further parameterize $\beta_i = \beta_{0i} + \log(N_i/10)$ where $N_i$ is the number of words in passage $i$ and $\beta_{0i}$ is the rescaled time intensity parameter in the scale of reading time per 10 words.

**Higher-order Model.** The latent ability $\theta$ and latent speed $\tau$ are further assumed to follow a bivariate normal distribution in the population:

$$\begin{pmatrix} \theta_j \\ \tau_j \end{pmatrix} \sim MVN \begin{pmatrix} \mu_\theta \\ \mu_\tau \end{pmatrix} \begin{pmatrix} \sigma_\theta^2 & \\ \sigma_{\theta\tau} & \sigma_\tau^2 \end{pmatrix} \tag{3}$$

The diagonal of the matrix indicates the variances of the latent variables. The off-diagonal of the matrix indicates the covariance between latent variables. For the model identification purpose, we constrain $\mu_\theta = \mu_\tau = 0$ and $\sigma_\theta^2 = 1$.

**Model-based ORF Scores.** Given the parameter estimates from the above joint model of accuracy and speed, we calculate the WCPM scores $s_j$ as a rate of accurate reading per minute. For person $j$ who read a total of $I$ passages, $s_j$ is

calculated as the expected number of words read correctly $E[U_{.j}]$ divided by the total reading time in seconds $E[T_{.j}]$ and further multiplied by 60:

$$s_j = \frac{E[U_{.j}]}{E[T_{.j}]} \times 60, \tag{4}$$

where

$$E[U_{.j}] = \sum_{i=1}^{I} N_i p_{ij} \tag{5}$$

and

$$E[T_{.j}] = \sum_{i=1}^{I} \exp(\beta_{0i} + \log(N_i/10) - \tau_j + \frac{1}{2\alpha_i^2}). \tag{6}$$

**ORF Score Equating**

Detailed descriptions of the four-step model-based equating method for ORF scores in a NEAT design can be found in Kara et al. (2023). In the current study, we follow the same steps and we focus on describing the estimation methods in the following paragraphs.

**Step 1. Concurrent Calibration.** In the equating scenario, we assume Group 1 read Form $V$ while the Group 2 read Form $U$ and both groups read some common external anchor passages given the NEAT design. We equate all passage parameters, including Form $V$, Form $U$ and anchor passages, by conducting concurrent calibration. We denote the response vector for an individual as $(\boldsymbol{u}, \boldsymbol{t})$. Further, each response vector consists of both observed and missing observations $(\boldsymbol{u}, \boldsymbol{t}) = (\boldsymbol{u}_{\text{obs}}, \boldsymbol{t}_{\text{obs}}, \boldsymbol{u}_{\text{miss}}, \boldsymbol{t}_{\text{miss}})$ because each person only read $K$ unique passages given his/her group membership and $C$ common passages. In this case, each pair of measurements in $(\boldsymbol{u}_{\text{miss}}, \boldsymbol{t}_{\text{miss}})$ is assumed to be missing completely at random and does not contribute to the complete data likelihood function. With $S = \{1, 2, ..., I\}$ denoting the passages read by a random person with $I = K + C$, the marginal likelihood function is given by

$$\mathcal{L} = \int_{\boldsymbol{\xi}} \left[ \prod_{i \in S} P(u_i; N_i, p_i(\theta)) f(t_i|\tau) \right] \phi_2(\boldsymbol{\xi}; \mu_{\boldsymbol{\xi}}, \boldsymbol{\Sigma}_{\boldsymbol{\xi}}) d\boldsymbol{\xi} \tag{7}$$

where $\boldsymbol{\xi} = (\theta, \tau)^\top \in \mathbb{R}^2$ and $\phi_2(\cdot; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes the bivariate normal density assumed for the distribution of latent variables.

To obtain the marginal maximum likelihood (MML) estimates of model parameters $\boldsymbol{\psi}$, including passage parameters (i.e., $\boldsymbol{a}, \boldsymbol{c}, \boldsymbol{\alpha}, \boldsymbol{\beta_0}$) and latent density parameters (i.e., $\sigma_\tau$, $\sigma_{\theta\tau}$), given the above marginal likelihood function, we adopt the Monte Carlo Expectation Maximization (MCEM) algorithm proposed in Potgieter, Kamata, and Kara (2017). This idea is similar to the EM algorithm (Dempster, Laird, & Rubin, 1977), where latent variables are treated as missing data and the expected complete data log-likelihood function conditional on observed random variables (E-step) is maximized in terms of model parameters (M-step). The iterative E-M process is repeated until model parameters converge. However, we adopt the Monte Carlo sampling approach for the E-step given that no closed form exists for the conditional expectations. Specifically, in the $k^{th}$ iteration, we implement $M_k$ independent draws from $\boldsymbol{\xi}|\boldsymbol{u}_{obs}, \boldsymbol{t}_{obs}, S, \hat{\boldsymbol{\psi}}_{k-1}$, the distribution of latent variable vector $\boldsymbol{\xi}$ conditional on observed observations and assumed true parameter values $\hat{\boldsymbol{\psi}}_{k-1}$ from the $k-1^{th}$ iteration. Details about the MCEM algorithm can be found in Potgieter et al. (2017). Due to the stochastic nature of MCEM, Wei and Tanner (1990) suggested to set small $M_k$ first and then large $M_k$ in the last several iterations to properly approximate the maximum likelihood solution. In this study, we implemented 200 iterations with $M_k = 1$ and then 5 iterations with $M_k = 10$ in all analyses.

Given the MML estimates $\hat{\boldsymbol{\psi}}_N$, we approximated the asymptotic error covariance matrix $\mathcal{I}_N^{-1}$ of these parameter estimates using a bootstrap approach. Specifically, we randomly drew 50 samples from the original data with replacement. Then, 50 sets of new parameter estimates for each of these 50 samples were obtained. The variance-covariance matrix for the 50 sets of these parameter estimates was calculated as an approximation of $\mathcal{I}_N^{-1}$. More specifically, the square root of the diagonal elements of $\mathcal{I}_N^{-1}$ is the $SE$s of the parameter estimates while the off-diagonal elements represent the covariances between parameters.

**Step 2. Scoring for Each Group.** Then, we consider expected a posteriori (EAP) scoring for each group separately based on $K$ unique passages they read given that common passages are external anchors in the current equating design. In the ideal

case where $\boldsymbol{\psi}$ are known, we make statistical inference of $\boldsymbol{\xi}$ for an individual with response vector $(\boldsymbol{u}, \boldsymbol{t})$ based on the following posterior:

$$f(\boldsymbol{\xi}|\boldsymbol{u}, \boldsymbol{t}, \boldsymbol{\psi}) = \frac{f(\boldsymbol{u}, \boldsymbol{t}|\boldsymbol{\xi}, \boldsymbol{\psi})f(\boldsymbol{\xi})}{\int f(\boldsymbol{u}, \boldsymbol{t}|\boldsymbol{\xi}, \boldsymbol{\psi})d\boldsymbol{\xi}} = \frac{f(\boldsymbol{u}, \boldsymbol{t}|\boldsymbol{\xi}, \boldsymbol{\psi})f(\boldsymbol{\xi})}{f(\boldsymbol{u}, \boldsymbol{t}|\boldsymbol{\psi})}. \tag{8}$$

The EAP estimator for $\boldsymbol{\xi}$ given $\boldsymbol{\psi}$ is an expectation over the posterior distribution in Equation 8:

$$\hat{\boldsymbol{\xi}}_{\psi} = \int_{\boldsymbol{\xi}} \boldsymbol{\xi} f(\boldsymbol{\xi}|\boldsymbol{u}, \boldsymbol{t}, \boldsymbol{\psi})d\boldsymbol{\xi} = \frac{1}{f(\boldsymbol{u}, \boldsymbol{t}|\boldsymbol{\psi})} \int_{\boldsymbol{\xi}} \boldsymbol{\xi} f(\boldsymbol{u}, \boldsymbol{t}|\boldsymbol{\xi}, \boldsymbol{\psi})f(\boldsymbol{\xi})d\boldsymbol{\xi} \tag{9}$$

The asymptotic $SE$ of $\hat{\theta}_{\psi}$ is the square root of the posterior variance:

$$V(\hat{\theta}_{\psi}) = \int_{\boldsymbol{\xi}} (\theta - \hat{\theta}_{\psi})^2 f(\boldsymbol{\xi}|\boldsymbol{u}, \boldsymbol{t}, \boldsymbol{\psi})d\boldsymbol{\xi} = \frac{1}{f(\boldsymbol{u}, \boldsymbol{t}|\boldsymbol{\psi})} \int_{\boldsymbol{\xi}} (\theta - \hat{\theta}_{\psi})^2 f(\boldsymbol{u}, \boldsymbol{t}|\boldsymbol{\xi}, \boldsymbol{\psi})f(\boldsymbol{\xi})d\boldsymbol{\xi} \tag{10}$$

Similarly, the asymptotic $SE$ of $\hat{\tau}_{\psi}$ is the square root of the posterior variance:

$$V(\hat{\tau}_{\psi}) = \int_{\boldsymbol{\xi}} (\tau - \hat{\tau}_{\psi})^2 f(\boldsymbol{\xi}|\boldsymbol{u}, \boldsymbol{t}, \boldsymbol{\psi})d\boldsymbol{\xi} = \frac{1}{f(\boldsymbol{u}, \boldsymbol{t}|\boldsymbol{\psi})} \int_{\boldsymbol{\xi}} (\tau - \hat{\tau}_{\psi})^2 f(\boldsymbol{u}, \boldsymbol{t}|\boldsymbol{\xi}, \boldsymbol{\psi})f(\boldsymbol{\xi})d\boldsymbol{\xi} \tag{11}$$

In real cases, $\boldsymbol{\psi}$ are most likely unknown. In this study, we consider two approaches to estimate EAP scores and their associated $SE$s when $\boldsymbol{\psi}$ are unknown: 1) the standard practice where the MML estimates $\hat{\boldsymbol{\psi}}_N$ are plugged in Equations 9 to 11 to replace $\boldsymbol{\psi}$ and 2) the alternative multiple imputation (MI) method proposed in Yang et al. (2012). The standard practice ignores the variability of $\hat{\boldsymbol{\psi}}_N$ as reflected by $\mathcal{I}_N^{-1}$. Thus, $\hat{\theta}_{\hat{\boldsymbol{\psi}}_N}$, $\hat{\tau}_{\hat{\boldsymbol{\psi}}_N}$, $V(\hat{\theta}_{\hat{\boldsymbol{\psi}}_N})$ and $V(\hat{\tau}_{\hat{\boldsymbol{\psi}}_N})$ ignore this variability and may be inaccurately estimated. This may further affect the estimation accuracy of $SE$s of equated ORF scores. Therefore, we adapt the MI method proposed in Yang et al. (2012) to our context. The formal justification of the MI procedure can be found in Yang et al. (2012). We delineate the MI algorithm to approximate $\hat{\theta}$, $\hat{\tau}$, $V(\hat{\theta})$ and $V(\hat{\tau})$ as follows:

1. $P$ sets of model parameter values are drawn from a multivariate normal distribution $\boldsymbol{\psi}_p \sim \text{MVN}(\hat{\boldsymbol{\psi}}_N, \mathcal{I}_N^{-1}), p = 1, 2, ...P$.

2. Plug each $\boldsymbol{\psi}_p$ into Equations 9 to 11 to calculate $\hat{\theta}_{\hat{\boldsymbol{\psi}}_p}$, $\hat{\tau}_{\hat{\boldsymbol{\psi}}_p}$, $V(\hat{\theta}_{\hat{\boldsymbol{\psi}}_p})$ and $V(\hat{\tau}_{\hat{\boldsymbol{\psi}}_p})$.

3. The MI EAP approximation to $\hat{\theta}$ and $\hat{\tau}$ is the empirical average, $\bar{\theta} \approx \frac{1}{P}\sum_{r=1}^{P} \hat{\theta}_{\boldsymbol{\psi}_p}$ and $\bar{\tau} \approx \frac{1}{P}\sum_{p=1}^{P} \hat{\tau}_{\boldsymbol{\psi}_p}$.

4. The MI variance approximation is $V(\bar{\theta}) \approx V_{\text{within}} + (1 + \frac{1}{P})V_{\text{between}}$, where $V_{\text{within}} = \frac{1}{P}\sum_{p=1}^{P} V(\hat{\theta}_{\psi_p})$ is an estimate of the within imputation variance, and $V_{\text{between}} = \frac{1}{P-1}\sum_{p=1}^{P}(\hat{\theta}_{\psi_p} - \bar{\theta})^2$ is an estimate of the between imputation variance. Same applies to $V(\bar{\tau})$.

**Step 3. Selecting Reference Passages.** In the current study, we focus on equating the ORF scores of Group 1 on Form $V$ to the scale of ORF scores on Form $U$ from Group 2. Thus, reference passages are Form $U$. We use the parameter estimates of these passages obtained from Step 1 to perform the equating.

**Step 4. Equating ORF Scores.** We consider the case where ORF scores from Group 1 are equated to Group 2. This is done by plugging passage parameter estimates $\hat{\boldsymbol{a}}$, $\hat{\boldsymbol{c}}$, $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\beta}}$ for Form $U$ and EAP scores, either based on the standard practice or the MI method, for Group 1 into Equation 4.

## Delta Method

Let $\boldsymbol{\gamma}$ denote the $(4K + 2) \times 1$ vector of parameters consisting of passage parameters $\boldsymbol{\lambda} = (\boldsymbol{a}^\top, \boldsymbol{c}^\top, \boldsymbol{\alpha}^\top, \boldsymbol{\beta_0}^\top)^\top$ for Form $U$ and $\boldsymbol{\xi} = (\theta, \tau)^\top$ for a random individual in Group 1. Let $s_e$ denote the equated ORF score for an individual which is a differentiable scalar function of $\boldsymbol{\gamma}$ same as that shown in Equation 4. Assuming that the estimator $\hat{\boldsymbol{\gamma}}$ fulfills $\sqrt{n}(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}) \xrightarrow{d} N(\mathbf{0}, \boldsymbol{\Sigma}_{\hat{\gamma}})$ as $n \to \infty$, as is the case for both the MML estimator and the EAP estimator, and $s_e{}'(\boldsymbol{\gamma}) \neq 0$, we can apply the delta method (Kendall, 1946) to obtain asymptotic $SE$ of $\hat{s}_e$. As $n \to \infty$,

$$\sqrt{n}(\hat{s}_e - s_e) \xrightarrow{d} N\Big[0, \Big(\frac{\partial s_e}{\partial \boldsymbol{\gamma}}\Big)^\top \boldsymbol{\Sigma}_{\hat{\gamma}} \frac{\partial s_e}{\partial \boldsymbol{\gamma}}\Big], \tag{12}$$

and the asymptotic $SE$ of $\hat{s}_e$ is the square root of:

$$V(\hat{s}_e) = \Big(\frac{\partial s_e}{\partial \boldsymbol{\gamma}}\Big)^\top \boldsymbol{\Sigma}_{\hat{\gamma}} \frac{\partial s_e}{\partial \boldsymbol{\gamma}}, \tag{13}$$

where $\frac{\partial s_e}{\partial \boldsymbol{\gamma}}$ is a $(4K + 2) \times 1$ vector consisting of partial derivatives of function $s_e$ with respect to passage parameters and latent variables; $\boldsymbol{\Sigma}_{\hat{\gamma}}$ is the asymptotic error variance-covariance matrix of these parameters. More specifically,

$\frac{\partial s_e}{\partial \boldsymbol{\gamma}} = (\frac{\partial s_e}{\partial \boldsymbol{a}}^\top, \frac{\partial s_e}{\partial \boldsymbol{c}}^\top, \frac{\partial s_e}{\partial \boldsymbol{\alpha}}^\top, \frac{\partial s_e}{\partial \boldsymbol{\beta_0}}^\top, \frac{\partial s_e}{\partial \theta}, \frac{\partial s_e}{\partial \tau})^\top$, where

$$\frac{\partial s_e}{\partial a_i} = \frac{60 \times N_i \phi(a_i\theta - c_i)\theta}{\sum_{i=1}^{K} \exp(\beta_{0i} + \log(N_i/10) - \tau_j + \frac{1}{2\alpha_i^2})} \tag{14}$$

$$\frac{\partial s_e}{\partial c_i} = -\frac{60 \times N_i \phi(a_i\theta - c_i)}{\sum_{i=1}^{K} \exp(\beta_{0i} + \log(N_i/10) - \tau_j + \frac{1}{2\alpha_i^2})} \tag{15}$$

$$\frac{\partial s_e}{\partial \alpha_i} = \frac{60 \times \sum_{i=1}^{K} N_i \Phi(a_i\theta - c_i) \times \exp(\beta_{0i} + \log(N_i/10) - \tau + \frac{1}{2\alpha_i{}^2})}{(\sum_{i=1}^{K} \exp(\beta_{0i} + \log(N_i/10) - \tau_j + \frac{1}{2\alpha_i^2}))^2 \alpha_i{}^3} \tag{16}$$

$$\frac{\partial s_e}{\partial \beta_{0i}} = -\frac{60 \times \sum_{i=1}^{K} N_i \Phi(a_i\theta - c_i) \times \exp(\beta_{0i} + \log(N_i/10) - \tau + \frac{1}{2\boldsymbol{\alpha}^2})}{(\sum_{i=1}^{K} \exp(\beta_{0i} + \log(N_i/10) - \tau_j + \frac{1}{2\alpha_i^2}))^2} \tag{17}$$

$$\frac{\partial s_e}{\partial \theta} = \frac{60 \times \sum_{i=1}^{K} N_i \phi(a_i\theta - c_i) a_i}{\sum_{i=1}^{K} \exp(\beta_{0i} + \log(N_i/10) - \tau_j + \frac{1}{2\alpha_i^2})} \tag{18}$$

$$\frac{\partial s_e}{\partial \tau} = \frac{60 \times \sum_{i=1}^{K} N_i \Phi(a_i\theta - c_i)}{\sum_{i=1}^{K} \exp(\beta_{0i} + \log(N_i/10) - \tau_j + \frac{1}{2\alpha_i^2})} \tag{19}$$

and

$$\boldsymbol{\Sigma}_{\hat{\gamma}} = \begin{pmatrix} \boldsymbol{\Sigma}_{\hat{\boldsymbol{\lambda}}} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_{\hat{\boldsymbol{\xi}}} \end{pmatrix}, \tag{20}$$

where $\boldsymbol{\Sigma}_{\hat{\boldsymbol{\lambda}}}$ and $\boldsymbol{\Sigma}_{\hat{\boldsymbol{\xi}}}$ can be obtained from the MML estimator and the EAP estimator, respectively; off-diagonal elements are 0s because the two estimators are independent. To investigate the impact of ignoring calibration error in the derivation of ORF equating *SE*s, we consider four scenarios as summarized in Table 1 in the specification of $\boldsymbol{\Sigma}_{\hat{\boldsymbol{\lambda}}}$ and $\boldsymbol{\Sigma}_{\hat{\boldsymbol{\xi}}}$ in Equation 20.

Table 1

*Considerations of Calibration Error*

| Scenario | $\boldsymbol{\Sigma}_{\hat{\boldsymbol{\lambda}}}$ | $\boldsymbol{\Sigma}_{\hat{\boldsymbol{\xi}}}$ | Abbreviation |
|---|---|---|---|
| 1. No Ignoration | $\boldsymbol{\Sigma}_{\hat{\boldsymbol{\lambda}}}$ | MI-based $V(\bar{\theta})$ and $V(\bar{\tau})$ | MI1 |
| 2. Partial Ignoration in $\boldsymbol{\Sigma}_{\hat{\boldsymbol{\lambda}}}$ | 0 | MI-based $V(\bar{\theta})$ and $V(\bar{\tau})$ | MI2 |
| 3. Partial Ignotation in $\boldsymbol{\Sigma}_{\hat{\boldsymbol{\xi}}}$ | $\boldsymbol{\Sigma}_{\hat{\boldsymbol{\lambda}}}$ | Standard practice $V(\hat{\theta})$ and $V(\hat{\tau})$ | Standard1 |
| 4. Complete Ignoration | 0 | Standard practice $V(\hat{\theta})$ and $V(\hat{\tau})$ | Standard2 |

*Note.* $\boldsymbol{\Sigma}_{\hat{\boldsymbol{\lambda}}}$ is the error covariance matrix of model parameters; $\boldsymbol{\Sigma}_{\hat{\boldsymbol{\xi}}}$ is the error covariance matrix of latent variables.

## 3. Simulation Study

We conducted a simulation study to evaluate the recovery of asymptotic $SE$s of equated ORF scores derived based on the delta method and to investigate the impact of ignoring calibration error in EAP scoring and/or delta method on equating $SE$s. We expect that asymptotic $SE$s of equated ORF scores should be more accurate and precise when calibration error is considered especially under large samples.

### Simulation Conditions

We manipulated two factors: sample size for each group $n$ as 100 (extremely small), 300 (small), 500 (median), or 1000 (large) and number of unique passages $K$ as 3 (short) or 6 (long). Factor levels were fully crossed and there were $3 \times 2 = 6$ simulation conditions. For each simulation condition, we calculated the equated ORF score $SE$s (as shown in Equation 13) in four ways as shown in Table 1. For the EAP estimators, numerical quadrature setup was the same with 150 nodes for each dimension, with $49^2 = 2401$ total function evaluations. The range of quadrature nodes was from -5 to 5. For the MI-based EAP estimator, we considered 20 imputations which was suggested as enough in Mislevy et al. (1994). We ran 500 replications for each condition.

### Data Generation

The data-generating model for all simulation conditions is the joint model of reading counts and response times described in Equations 1 to 3. True passage parameter values for Form $V$ and Form $U$ can be found in Table 2. We generated nonequivalent groups by drawing $\theta$ and $\tau$ from two bivariate normal distributions (as indicated in Equation 3) with different mean vectors but the same variance-covariance matrix. For Group 1, $\mu_{\theta 1} = 0.5$ and $\mu_{\tau 1} = 0.2$, while for Group 2, $\mu_{\theta 2} = -0.5$ and $\mu_{\tau 2} = -0.2$. That is, Group 1 had higher latent ability and speed levels than Group 2. For both groups, $\sigma_\tau^2 = 0.16$ and $\sigma_{\theta\tau} = 0.16$.

Table 2

*True Passage Parameters*

| | $a$ | $c$ | $\alpha$ | $\beta_0$ | $N$ |
|---|---|---|---|---|---|
| Form $V$ | 0.500 | -1.384 | 6.176 | 1.723 | 47 |
| | 0.444 | -1.299 | 5.747 | 1.768 | 49 |
| | 0.557 | -1.537 | 4.841 | 1.711 | 50 |
| | 0.505 | -1.442 | 5.064 | 1.750 | 50 |
| | 0.447 | -1.318 | 4.408 | 1.705 | 50 |
| | 0.476 | -1.370 | 3.936 | 1.752 | 49 |
| Form $U$ | 0.569 | -1.291 | 5.808 | 1.963 | 47 |
| | 0.569 | -1.348 | 4.559 | 1.972 | 54 |
| | 0.603 | -1.250 | 5.566 | 1.879 | 50 |
| | 0.618 | -1.338 | 4.239 | 1.863 | 49 |
| | 0.571 | -1.316 | 3.795 | 1.894 | 49 |
| | 0.588 | -1.206 | 5.248 | 1.916 | 50 |
| Anchor | 0.575 | -1.394 | 6.821 | 1.806 | 54 |

*Note.* First three passages in Form $V$ and Form $U$ were used under simulation condition $K = 3$.

**Evaluation Criteria**

To assess the recovery of asymptotic $SE$s of equated ORF scores, we considered raw bias and root mean squared error (RMSE) as outcome measures. Specifically, the raw bias and RMSE are calculated as:

$$\text{Raw Bias} = \frac{1}{R}\sum_{r=1}^{R} SE(\hat{s}_{e_r}) - SE(s_e), \tag{21}$$

$$\text{RMSE} = \sqrt{\frac{1}{R}\sum_{r=1}^{R}(SE(\hat{s}_{e_r}) - SE(s_e))^2}, \tag{22}$$

where $SE(\hat{s}_{e_r})$ is the $SE$ estimate (standard or MI-based) for equated ORF score in replication $r$; $R$ is total number of replications; $SE(s_e)$ is the criterion $SE$ value. We used the empirical standard deviation (ESD) of $\hat{s}_{e_r}$ across replications as the criterion

$SE$ in each simulation condition. Specifically, we used MI-based $\bar{\theta}$ and $\bar{\tau}$ calculate criterion $SE$s. We expect that these are closer to the true equated ORF score $SE$s than those calculated from standard EAP scores that ignore calibration errors. We used custom codes in $R$ language (R Development Core Team, 2018). All computations were run on a high-performance cluster featuring the CentOS 7 operating system.

## 4. Results

In this section, we summarize simulation results to present the recovery of equated ORF score $SE$s derived based on delta method. More specifically, we aim to compare the recovery of equated ORF score $SE$s according to four considerations of calibration error in the implementation of delta method as shown in Table 1. We refer to these four approaches using their abbreviations shown in Table 1, i.e., MI1, MI2, Standard1 and Standard2 in following paragraphs, figures and tables.

Figure 1 shows the distributions of equated ORF score $SE$s under all simulated conditions using box plots. Box plots exhibit five important characteristics of the distributions: minimum, maximum, median, first and third quartiles. Under each simulation condition, the five box plots summarize the distribution of criterion $SE$s (ESD), average $SE$s across replications based on approaches Standard1, Standard2, MI1 and MI2, respectively.

In general, all four approaches yielded comparable equated ORF score $SE$s as the criterion $SE$s especially under large samples. This indicates the accuracy of our derivation of the asymptotic $SE$s of ORF score equating via the delta method. The discrepancies of estimated $SE$ distributions, on the other hand, stemmed from different considerations of calibration error in the computation of equated ORF score $SE$s. Specifically, when calibration error was considered in both latent variable scoring and the delta method (MI1), estimated $SE$s were larger than the other three approaches where calibration error was ignored in either or both steps. Further, we found that more accurate equated ORF score $SE$s can be found under larger sample size and longer test length. In fact, when $n = 500$ and $K = 6$ or $n = 1000$ and $K = 6$, we

observed a clear pattern that MI1 yielded comparable $SEs$ as the criterion $SE$s, while Standard1, Standard2 and MI2 underestimated equated ORF score $SE$s given all five distributional characteristics shown in the box plot. This is as expected given that the delta method we employed based on the asymptotic theory and is supposed to yield accurate $SE$ estimates under large samples.

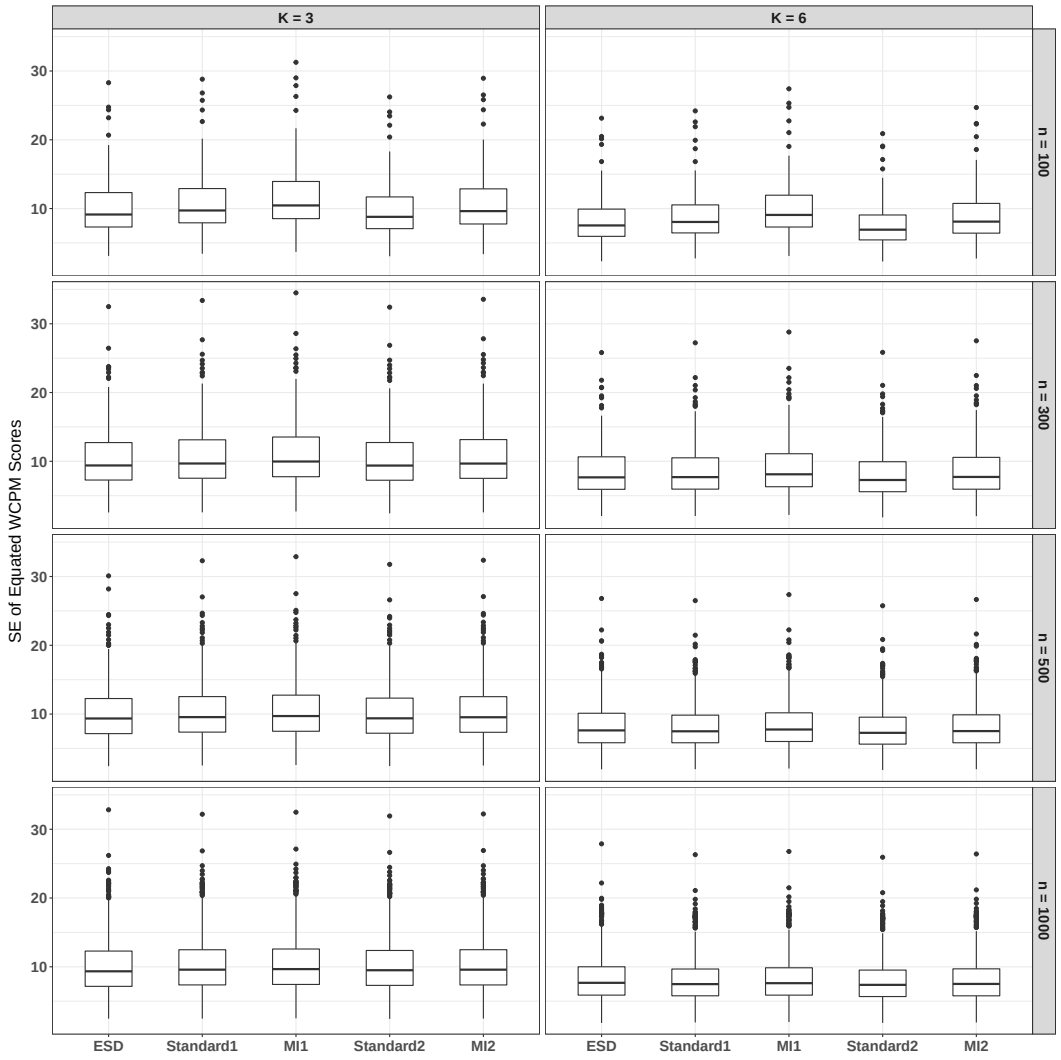


*Figure 1*. Summary of $SE$s under all simulation conditions.

*Note.* $K$ = number of unique passages; $n$ = group size; ESD = Empirical standard deviation of WCPM scores.

Figure 2 shows the distribution of raw bias of equated ORF score $SE$s under all simulated conditions using box plots. In general, distributions close to and center around 0 indicate satisfactory recovery of equating $SE$s. Similar to the finding based on Figure 1, $SE$s based on the MI1 approach had higher accuracy when sample size larger

and test length longer, e.g., $n = 1000$ and $K = 6$.

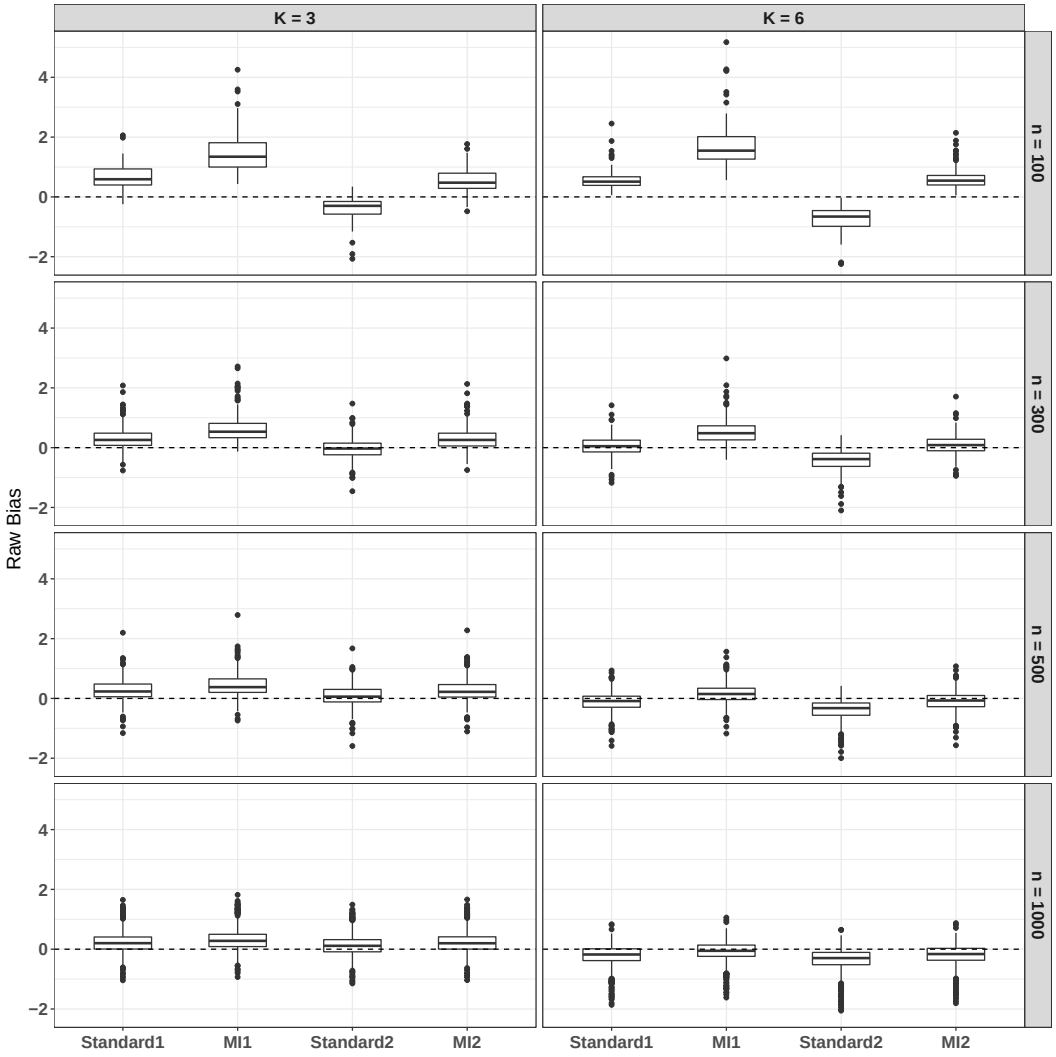


*Figure 2*. Raw bias of $SE$s under all simulation conditions.

*Note.* $K$ = number of unique passages; $n$ = group size; ESD = Empirical standard deviation of WCPM scores.

Table 3 shows the RMSE of the equated ORF score $SE$s obtained from the four approaches: Standard1, Standard2, MI1 and M2. Given the scale of equated ORF score $SE$s, both approaches yielded adequate recovery based on the calculated values of all three outcome measures. More specifically, MI1 had smaller average RMSE under conditions with larger sample size and longer test length $n = 500, K = 6$ and $n = 1000, K = 6$. This indicates accurate and precise equating $SE$ estimates obtained by the MI1 approach especially when sample size and test length increase.

Table 3

*Average RMSE Values*

| | | Standard1 | MI1 | Standard2 | MI2 |
|---|---|---|---|---|---|
| $n = 100$ | $K = 3$ | 1.653 | 2.142 | **1.458** | 1.483 |
| | $K = 6$ | 1.982 | 2.349 | 2.228 | **1.765** |
| $n = 300$ | $K = 3$ | 1.383 | 1.473 | **1.340** | 1.351 |
| | $K = 6$ | 2.002 | 1.931 | 2.131 | **1.918** |
| $n = 500$ | $K = 3$ | 1.324 | 1.358 | **1.295** | 1.304 |
| | $K = 6$ | 1.999 | **1.920** | 2.080 | 1.955 |
| $n = 1000$ | $K = 3$ | 1.311 | 1.321 | **1.296** | 1.300 |
| | $K = 6$ | 2.045 | **1.991** | 2.088 | 2.022 |

*Note.* $K$ = number of unique passages; $n$ = group size; Standard = standard practice that ignores calibration error in scoring; MI = multiple imputation approach that accounts for calibration error in scoring. Smaller values are boldface.

## 5. Discussion

The current study provides an analytical derivation of asymptotic $SE$s of ORF score equating based on the delta method. Same as other assessment scenarios, reporting equating $SE$s is advocated in the usage of ORF assessment. We conducted a simulation study that demonstrated the accuracy of our derivation given satisfactory recovery of equated ORF score $SE$s. The model-based equating method we implemented in the current study consisted of concurrent calibration and ORF score equating. Ignoring calibration error, although the standard practice in IRT scoring, may underestimate asymptotic $SE$s of ORF score equating. Therefore, the simulation study also examined the impact of adopting either the standard practice or the multiple imputation method proposed in Yang et al. (2012) that takes into account calibration error in ORF scoring on recovery of equated ORF score $SE$s. Results indicated that ignoring calibration error may yield larger bias and variability in equated ORF score $SE$s especially under large sample size and long test length.

There are several limitations in the current study. First, given limited computation resource, the number of replication was 500 in the simulation study. More accurate criterion $SE$s can be obtained by using larger number of replications (e.g., 10,000). Second, although the derivation of asymptotic $SE$s of ORF score equating can be generalized to different equating designs, it is limited to the specific parametric model used in the current study. For example, count data model can be beta-binomial instead of binomial (Qiao, Kamata, Kara, Potgieter, & Nese, submitted), in which case the derivation should be modified accordingly. Lastly, the derived $SE$s of ORF score equating are based on asymptotic theory. Therefore, applications are suggested to have large sample size and long test length.

In practice, a researcher's choice of calibration and scoring methods used for the ORF model may depend on several factors such as computational resources, sample size and test length. Given that number of passage is usually small, we suggest the use of EAP scores. Further, although the MI approach may be inaccessible in practice, we emphasize that equated ORF score $SE$s may be underestimated if calibration error is ignored.

References

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. (2018). *Standards for educational and psychological testing*. American Educational Research Association.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, *39*(1), 1–22.

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.

Kara, Y., Kamata, A., Qiao, X., Potgieter, C. J., & Nese, J. F. (2023). Equating oral reading fluency scores: A model-based approach. *Educational and Psychological Measurement*, 00131644221148122.

Kendall, M. G. (1946). The advanced theory of statistics. *The advanced theory of statistics.*(2nd Ed).

Mislevy, R. J., Wingersky, M. S., & Sheehan, K. M. (1994). Dealing with uncertainty about item parameters: Expected response functions. *ETS Research Report Series*, *1994*(1), i–20.

Ogasawara, H. (2000). Asymptotic standard errors of irt equating coefficients using moments. *Economic Review (Otaru University of Commerce)*, *51*(1), 1–23.

Ogasawara, H. (2001a). Item response theory true score equatings and their standard errors. *Journal of Educational and Behavioral Statistics*, *26*(1), 31–50.

Ogasawara, H. (2001b). Standard errors of item response theory equating/linking by response function methods. *Applied Psychological Measurement*, *25*(1), 53–67.

Potgieter, C. J., Kamata, A., & Kara, Y. (2017). An em algorithm for estimating an oral reading speed and accuracy model. *arXiv preprint arXiv:1705.10446*.

Qiao, X., Kamata, A., Kara, Y., Potgieter, C. J., & Nese, J. F. (submitted). *Applied Psychological Measurement*.

R Development Core Team. (2018). *R: A language and environment for statistical*

*computing* [Computer Software]. Vienna, Austria: R Foundation for Statistical Computing, ISBN 3-900051-07-0.

van der Linden, W. J. (2006). A lognormal model for response times on test items. *Journal of Educational and Behavioral Statistics*, *31*(2), 181–204.

Wei, G. C., & Tanner, M. A. (1990). A monte carlo implementation of the em algorithm and the poor man's data augmentation algorithms. *Journal of the American statistical Association*, *85*(411), 699–704.

Yang, J. S., Hansen, M., & Cai, L. (2012). Characterizing sources of uncertainty in item response theory scale scores. *Educational and psychological measurement*, *72*(2), 264–290.