Item Calibration Methods with Multiple Sub-scale Multistage Testing

Chun Wang
University of Washington
Ping Chen
Beijing Normal University
Shengyu Jiang
University of Minnesota

Correspondence concerning this manuscript should be addressed to Chun Wang at:

312E Miller Hall
Measurement and Statistics
College of Education, University of Washington
2012 Skagit Ln, Seattle, WA 98105
e-mail: wang4066@uw.edu

Acknowledgement: This research was supported by the IES R305D170042 (R305D160010) and NSF SES-165932. The authors would like to especially thank Drs. Yue Helena Jia, Pual Jewsbury, and Meng Wu for consolidating the research idea, Dr. Jing Chen from NCES for agreeing to share the data, and David Freund for preparing the real data.

Citation: Wang, C., Chen, P., & Jiang, S. (2019). Item calibration methods with multiple subscale multistage testing. *Journal of Educational Measurement*. https://doi.org/10.1111/jedm.12241
Related code can be downloaded at: https://sites.uw.edu/pmetrics/publications-and-source-code/

Item Calibration Methods with Multiple Subscale Multistage Testing

Abstract

Many large-scale educational surveys have moved from linear form design to multistage testing (MST) design. One advantage of MST is that it can provide more accurate latent trait (θ) estimates using fewer items than required by linear tests. However, MST generates incomplete response data by design; hence questions remain as to how to calibrate items using the incomplete data from MST design. Further complication arises when there are multiple correlated subscales per test, and when items from different subscales need to be calibrated according to their respective score reporting metric. The current calibration-per-subscale method produced biased item parameters, and there is no available method for resolving the challenge. Deriving from the missing data principle, we showed when calibrating all items together, the Rubin's (1976) ignorability assumption is satisfied such that the traditional single-group calibration is sufficient. When calibrating items per subscale, we proposed a simple modification to the current calibration-per-subscale method that helps reinstate the missing-at-random assumption and therefore corrects for the estimation bias that is otherwise existent. Three mainstream calibration methods are discussed in the context of MST, they are the marginal maximum likelihood estimation (MML), the expectation maximization (EM) method, and the fixed parameter calibration (FPC). An extensive simulation study is conducted and a real data example from NAEP is analyzed to provide convincing empirical evidence.

Key words: multistage testing, missing data, marginal maximum likelihood, EM

1. Introduction

With the advent of web-based technology, computer based testing (a.k.a., online testing) is becoming the mainstream form of large-scale educational assessments. The landscape of educational assessment is changing rapidly with the growth of computer-administered tests. As an example, National Assessment of Educational Progress (NAEP), the "*largest nationally representative and continuing assessment*" (e.g., Beaton & Zwick, 1992), has moved from paper based assessment (PBA) to digitally based assessment (DBA) recently.

A particular mode of DBA that NAEP has piloted for Mathematics is the multistage testing (MST), which refers to a testing format where "subsets of test items are presented to students based on item difficulty and student performance" (Governing Board and NAEP Resources[1]). Figure 1 illustrates a simple, two-stage MST design. The routing block contains items spread across a typical range of difficulty levels in PBA, and the targeted blocks differ by difficulty—blocks of easy, medium, and hard items.

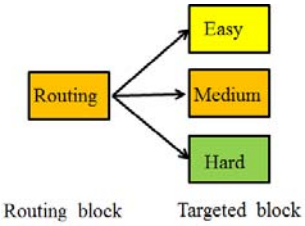

Figure 1. An illustration of a two-stage MST design used in NAEP

Compared to the linear form design, the MST design has a profound advantage. That is, due to length constraints and the demands placed on the single set of items, linear form tests may provide little information to certain subgroups (mostly highly achieving subgroups or low achieving subgroups) because there are not enough items with appropriate difficulty levels to measure students in those subgroups. A MST, however, tailors the set of items (i.e., target block)

[1] Retrieved from https://www.nagb.org/content/nagb/assets/documents/what-we-do/quarterly-board-meeting-materials/2014-11/tab02-governing-board-and-naep-resources.pdf

a student sees to the student's individual ability level, so that no student receives too many overly easy or difficult items. Consequently, MST can provide more accurate latent trait (θ) estimates using fewer items than required by PBA (e.g., Weiss, 1982; Wainer, 1990). Moreover, the computer-based nature of MST yields many other advantages, such as new item formats, new types of skills that can be measured, easier and faster data analysis, and richer behavior data collection such as item response time (as part of behavior/process data) (e.g., Wang, Zheng, & Chang, 2014).

Despite of the advantages, MST generates incomplete response data by design; hence questions remain as to whether the item calibration procedure for the traditional linear forms (e.g., Mislevy, 1991; Mislevy, Beaton, Kaplan, & Sheehan, 1992) can still apply. Widely used calibration methods for linear forms include the marginal maximum likelihood estimation with expectation maximization implementation (MMLE/EM; Bock & Aitkin, 1981), the expectation maximization (EM) algorithm (Dempster, Laird, & Rubin, 1977; Woodruff & Hanson,1996), and the fixed parameter calibration (FPC) methods (Ban, Hanson, Wang, Yi, & Harris, 2001; Chen & Wang, 2016, Chen, Wang, Xin, & Chang, 2017; Kim, 2006). A default assumption made by all three methods is that the sample is drawn from a single population, although the multiple group versions of all three methods have also been developed (Lissitz, Jiao, Li, et al., 2014). While MMLE/EM often assumes θ follows a normal distribution, both EM and FPC allow more flexible θ distributions.

Given the MST design in Figure 1, students routed to each module in the second stage naturally form three separate subgroups, whose θ distributions differ. The three non-equivalent groups share the same routing block, which serves as the linkage to put all items on the same scale. In this regard, it seems intuitively reasonable to assume that there are multiple subgroups

and the subgroup structure should be taken into account during the calibration procedure. Indeed, several recent studies (Cai, Roussos, & Wang, 2018; Lu, Jia, Wu, 2017) have explored the multiple-group MML method for MST item calibration. Their results showed that the multiple-group MML performed poorly, yielding large item parameter bias, whereas the single-group MML performed well. However, not clear reason is provided to explain the results.

In addition, another layer of complexity arises when the assessment covers multiple content subdomains. For instance, the mathematics assessment in NAEP has five subscales, "Number properties and operations", "Measurement", "Geometry", "Data analysis, Statistics, and Probability", and "Algebra". For score reporting purposes, items from each subscale need to be calibrated on their respective scale. Traditionally, the item calibration on each subscale is conducted separately using the unidimensional item response theory (IRT) models, and then a composite score, which is a weighted combination of the subscale scores[2], is created to report the overall mathematics performance. However, this calibration-per-subscale approach failed to recover item parameters properly within the MST design (e.g., Lu, Jia, & Wu, 2018; Wu & Lu, 2017; Wu & Xi, 2017), and no viable alternative was provided.

To sum up, there are two scenarios where MST item calibration has been explored: the first one is when all items are put on a single unidimensional scale, and the second one is when items from different content subdomains are put on separate unidimensional scales. The aim of the paper is two-fold: (1) to provide reasons why the current MST item calibration approaches are unsuccessful, including the multiple-group MML (Cai, et al., 2018; Lu, et al., 2017) for the first scenario and the single-group calibration-per-subscale for the second scenario (Wu & Xi, 2017; Wu & Lu, 2017); and (2) to propose a new method that resolves the challenge in the

[2] https://nces.ed.gov/nationsreportcard/tdw/analysis/trans.aspx

second scenario. The proposed solution is grounded in Rubin (1976)'s missing data theory, which provides a streamlined framework to explore the MST item calibration for both scenarios. Please note that the second scenario was motivated from the operational NAEP analysis, but the solutions provided could be apply to other operational designs similar to NAEP.

The rest of the paper is organized as follows. We first briefly introduce the unidimensional two-parameter logistic (2PL) model as the underlying IRT models throughout the study. Then we will describe the three commonly used item calibration methods: MMLE/EM, EM, and FPC. All these methods could be used with the MST data. In the next section, we will introduce Rubin (1976)'s missing data theory and its application to the MST design. In particular, we will explain why the current calibration-per-subscale method with the MST design is inadequate, and present a new, simple solution. Two simulation studies are presented, followed by a real data illustration. A discussion is presented in the end.

2. Models

The unidimensional 2PL model is used throughout the paper. For 2PL, the item response function for item *j* takes the following form

$$P_j(\theta) = \frac{e^{1.7a_j(\theta - b_j)}}{1 + e^{1.7a_j(\theta - b_j)}}, \quad (1)$$

where subscript j indicates item. a_j and b_j denote item discrimination and difficulty parameters respectively, and θ denotes the latent trait measured by the test. Here "1.7" is a scaling factor to equate the logistic form with the normal ogive form.

3. Existing Item Calibration Methods

3.1 Marginal Maximum Likelihood Estimation/Expectation Maximization (MMLE/EM)

The MMLE/EM algorithm (or MML for short hereafter) for IRT parameter estimation when the response data is complete has been well established in the literature (e.g., Bock & Aitkin, 1981; Mislevy, 1984). Suppose a *J*-item test is given to *N* examinees, resulting in an *N*-by-*J* binary response matrix $\boldsymbol{Y}$. Assuming all items are modelled by 2PL, and let $\boldsymbol{\Delta} = (\boldsymbol{a}, \boldsymbol{b})$ denote the set of unknown item parameters, which are the target parameters in item calibration. The joint likelihood can be easily written as

$$L(\boldsymbol{\Delta}, \boldsymbol{\theta}|\boldsymbol{Y}) = \prod_{i=1}^{N} \prod_{j=1}^{J} \left[P_j(\theta_i)^{y_{ij}} (1 - P_j(\theta_i))^{1-y_{ij}}\right], \quad (2)$$

due to the local independence assumption. Let $P(\boldsymbol{y}_i|\theta_i, \boldsymbol{\Delta}) = \prod_{j=1}^{J} \left[P_j(\theta_i)^{y_{ij}} (1 - P_j(\theta_i))^{1-y_{ij}}\right]$ denote the joint probability of $\boldsymbol{y}_i$ for notational simplicity. Then the marginal likelihood of $\boldsymbol{\Delta}$ is

$$L(\boldsymbol{\Delta}|\boldsymbol{Y}) = \prod_{i=1}^{N} \int P(\boldsymbol{y}_i|\theta_i, \boldsymbol{\Delta})\, g\left(\theta \middle| \mu_\theta, \sigma_\theta^2\right) d\theta, \quad (3)$$

where $g\left(\theta \middle| \mu_\theta, \sigma_\theta^2\right)$ denotes the density function of θ in the population, and μ_θ and σ_θ^2 are its mean and variance respectively. Here in Eq. (3), it is assumed that there is one population from which the sample is drawn. However, the MML method could also be generalized to multiple group scenario such that the population mean and variance will be group specific (e.g., Mislevy, et al., 1992; Cai, Yang, & Hansen, 2011).

To remove the scale indeterminacy inherent in the IRT models, in Eq. (3), one often assumes that the latent trait θ follows a standard normal distribution (i.e., $g\left(\theta \middle| \mu_\theta = 0, \sigma_\theta^2 = 1\right)$). The marginal likelihood in Eq. (3) cannot be directly maximized easily because there is no closed form solution of $\boldsymbol{\Delta}$, and finding numerical solution means searching in a $2\times J$-dimensional space. The EM algorithm, however, provides a viable computational tool to simplify the direct maximization of the marginal likelihood (Bock & Aitkin, 1981).

In essence, the EM algorithm alternates between the E-step and M-step. In the E-step, the conditional expectation of the complete data log-likelihood (i.e., $l(\boldsymbol{\Delta}|\boldsymbol{Y},\boldsymbol{\theta}) = \log(L(\boldsymbol{\Delta}|\boldsymbol{Y},\boldsymbol{\theta}))$) with respect to the missing data (in this case, θ) is obtained, denoted as

$$E_{\boldsymbol{\theta}|\boldsymbol{Y},\boldsymbol{\Delta}^{r}}\big(l(\boldsymbol{\Delta}|\boldsymbol{Y},\boldsymbol{\theta})\big) = E_{\boldsymbol{\theta}|\boldsymbol{Y},\boldsymbol{\Delta}^{r}}\big(\log\big(L(\boldsymbol{\Delta}|\boldsymbol{Y},\boldsymbol{\theta})\big)\big), \tag{4}$$

where $\boldsymbol{\Delta}^{\boldsymbol{r}}$ denote the parameter estimates from the rth iteration. The notation $E_{\boldsymbol{\theta}|\boldsymbol{Y},\boldsymbol{\Delta}^{r}}$ implies that the expectation is taken with respect to the conditional distribution of θ (i.e., missing data) given the observed data ($\boldsymbol{Y}$) and provisional parameter estimates, $P\big(\theta\big|\boldsymbol{Y},\boldsymbol{\Delta}^{\boldsymbol{r}}\big)$. This conditional expectation is maximized to obtain the MLE of $\boldsymbol{\Delta}$ in the M-step. This way, the $2\times J$-dimensional maximization challenge is reduced to searching a numerical solution in a 2-dimensional space, which is much more feasible.

3.2 Expectation Maximization (EM) algorithm

While the above MML method treats the EM algorithm as a tool to reduce the computational complexity of directly maximizing the marginal likelihood, the item calibration can also proceed directly from the principal idea of the EM algorithm (Bock & Aitkin, 1981; Dempster et al., 1977; Rubin, 1991; Rubin & Thayer, 1982). In this case, the unknown latent trait θ is considered "missing" data. To model the distribution of θ flexibly, let us consider the discrete values θ_k $(k = 1, \dots, K)$ and their associated unknown probabilities π_k $(k = 1, \dots, K)$ (Kim, 2006). Here K is the total number of quadrature points along the θ continuum. Under this assumption, θ distribution can be recovered via the probability mass function π_k where $\sum_{k=1}^{K}\pi_k = 1$. In this regard, both the item parameters $\boldsymbol{\Delta} = (\boldsymbol{a},\boldsymbol{b})$ and $\boldsymbol{\pi} = (\pi_1, \dots, \pi_K)$ are unknown parameters. The IRT latent scale can be fixed by setting the mean of θ at 0, i.e., $\sum_{k=1}^{K}\pi_k\theta_k = 0$, and by setting its variance at 1.

The EM algorithm again proceeds by alternating between the E-step and the M-step. Here, the conditional expectation is slightly different from Eq. (4) as follows,

$$E_{\boldsymbol{\theta}|\boldsymbol{Y},\boldsymbol{\Delta}^r,\boldsymbol{\pi}^r}\left(\log\left(L(\boldsymbol{\Delta},\boldsymbol{\pi}|\boldsymbol{Y},\boldsymbol{\theta})\right)\right) = E_{\boldsymbol{\theta}|\boldsymbol{Y},\boldsymbol{\Delta}^r,\boldsymbol{\pi}^r}\left(\log\prod\prod_{i=1}^{N} L(\boldsymbol{\Delta},\boldsymbol{\pi}|\boldsymbol{Y}_i,\theta_i)\right)$$

$$= \sum_{i=1}^{N}\int \log L(\boldsymbol{\Delta},\boldsymbol{\pi}|\boldsymbol{Y}_i,\theta_i)\, P\left(\theta\middle|\boldsymbol{Y}_i,\boldsymbol{\Delta}^{\boldsymbol{r}},\boldsymbol{\pi}^r\right)d\theta$$

$$\propto \sum_{i=1}^{N}\sum_{k=1}^{K} \log(L(\boldsymbol{\Delta}|\boldsymbol{Y}_i)\pi_k) \times P\left(\theta_k\middle|\boldsymbol{Y}_i,\boldsymbol{\Delta}^{\boldsymbol{r}},\boldsymbol{\pi}^r\right), \qquad (5)$$

where $P\left(\theta_k\middle|\boldsymbol{Y}_i,\boldsymbol{\Delta}^{\boldsymbol{r}},\boldsymbol{\pi}^r\right)$ is the posterior distribution of θ_k given $\boldsymbol{Y}_i$, $\boldsymbol{\Delta}^{\boldsymbol{r}}$ and $\boldsymbol{\pi}^r$. Then in the M-step, the conditional expectation is maximized with respect to both $\boldsymbol{\Delta}$ and $\boldsymbol{\pi}$. Solving for item parameters remains the same as in section 3.1, whereas π_k^r is updated via a simple, closed-form solution as follows

$$\pi_k^{r+1} = \frac{f_k^r}{\sum_{k=1}^{K} f_k^r} = \frac{f_k^r}{N}, \qquad (6)$$

where $f_k^r = \sum_{i=1}^{N} P(\theta_k|\boldsymbol{Y}_i,\boldsymbol{\Delta}^{\boldsymbol{r}},\boldsymbol{\pi}^r)$. Within each EM cycle, to fix the latent scale, a few standardization steps need to be in place. In particular, let $\mu^{r+1} = \sum_{k=1}^{K}\pi_k^{r+1}\theta_k^{r+1}$ and $\sigma^{2,r+1} = \sum_{k=1}^{K}\pi_k^{r+1}(\theta_k^{r+1}-\mu^{r+1})^2$ be the provisional mean and variance of θ, then the discrete quadrature points are standardized by updating θ_k^{r+1} with $\frac{\theta_k^{r+1}-\mu^{r+1}}{\sigma^{r+1}}$. Accordingly, the provisional item parameter estimates are updated as follows: a^{r+1} is updated with $a^{r+1}\times\sigma^{r+1}$, b^{r+1} is updated with $\frac{b^{r+1}-\mu^{r+1}}{\sigma^{r+1}}$, and π_k^{r+1} is updated with $\pi_k^{r+1}\times\sigma^{r+1}$.

Although in the above exposition, $\boldsymbol{\pi}$ is estimated for a single θ distribution, the EM algorithm can also be extended for multiple group calibration. That is, group specific $\boldsymbol{\pi}$'s could be estimated for each subpopulation separately. One advantage of the EM algorithm compared to MML is that the distribution of θ does not have to be specified in advance, and hence it is more

flexible to deal with non-normal θ distributions. This is especially desirable in the multiple group calibration approach when the group specific θ distributions are unknown.

3.3 Fixed Parameter Calibration (FPC)

Fixed parameter calibration refers to fixing a subset of item parameters at their previously estimated values and calibrating the remaining items so that their item parameters are placed on the same, fixed scale. In this case, the scale of θ is naturally determined via the fixed parameters, and hence no constraints need to be added. This method is often used in online calibration scenario where new items are calibrated while holding the operational item parameters as fixed (e.g., Chen & Wang, 2016; Kim, 2006). Both the aforementioned MML and EM methods can be used in FPC. With the former method, if the θ distribution is assumed normal, then its mean and variance can be freely estimated; whereas with the latter method, the standardization steps are no longer needed. For more details, please refer to Chen et al. (2017) or Kim (2006). In this paper, will consider the EM algorithm coupled with FPC such that the θ distribution does not have to be pre-specified. Moreover, FPC can also be used with both single group and multiple group calibration approaches (Kim & Kolen, 2016).

4. Item Calibration with Missing Data

By nature, the multistage testing generates incomplete response data because after the routing stage, each examinee is routed to one module in the remaining stages that matches closely with his/her ability level. Mislevy and Sheenan (1989) showed that in incomplete designs, the use of MML could be justified from Rubin's (1976) general theory on inference in the presence of missing data. In particular, Mislevy and Wu (1996) argued that missing data due to MST (or adaptive) testing can be ignored when making inference about θ because the chance

for an item to be missing depends on observed responses from previous items but not on unobserved responses. However, they did not discuss the impact of missing data on item calibration. Eggen and Verhelst (2011) first provided a brief justification of using MML in the MST item calibration, but they did not mention the scenario when the test contains multiple subscales. In this section, we intend to provide a comprehensive discussion with regard to the missing data mechanism of the MST design within Rubin' (1976) framework, especially the implications of missing data on item calibration when the test contains multiple subscales. Please note that for exposition simplicity, we assume there is only one form per module in the MST design in this paper. However, in practice, there is oftentimes multiple, parallel forms per module. Because the parallel forms are usually randomly assigned, the missing data resulting from this random assignment is completely random and hence it can be ignored.

Essential to Rubin's (1976) theory is the stochastic nature of the missing data mechanism (Little & Rubin, 1987), denoted as

$$h_{\varphi}(\boldsymbol{M} = \boldsymbol{m}|\boldsymbol{Y} = \boldsymbol{y}), \tag{7}$$

where $M = \left(M_1, \ldots, M_J\right)$ is the missing data indicator, indicating whether Y_j is actually observed (i.e., $m_j = 1$) or missing (i.e., $m_j = 0$). y_j is the response on item j. Eq. (7) defines the process that causes the missing data, with the parameter φ that governs the missing mechanism.

In the incomplete design, we have a sample realization of $\boldsymbol{M}$ and $\boldsymbol{Y}_{obs}$ ("obs" denotes observed responses). So we can only estimate the item parameters of interest (i.e., $\boldsymbol{\Delta}$) based on partially observed $\boldsymbol{Y}$, which is the marginal joint distribution of $\boldsymbol{M}$ and $\boldsymbol{Y}_{obs}$ as,

$$\int_{\boldsymbol{y}_{mis}} f_{\varphi}(\boldsymbol{m}, \boldsymbol{y}|\boldsymbol{\Delta})\, d\boldsymbol{y}_{mis} = \int_{\boldsymbol{y}_{mis}} f(\boldsymbol{y}|\boldsymbol{\Delta}) h_{\varphi}(\boldsymbol{m}|\boldsymbol{y})\, d\boldsymbol{y}_{mis} \tag{8}$$

where $f_{\varphi}(\boldsymbol{m}, \boldsymbol{y}|\boldsymbol{\Delta})$ is the joint distribution of the complete data (i.e., $\boldsymbol{Y} = (\boldsymbol{Y}_{obs}, \boldsymbol{Y}_{mis})$) and the missing indicators. According to Rubin (1976), if the process that causes missing data can be

ignored, then Eq. (8) is equivalent to $\int_{\boldsymbol{y}_{mis}} f_{\varphi}(\boldsymbol{m}, \boldsymbol{y}|\boldsymbol{\Delta})\, d\boldsymbol{y}_{mis} = f(\boldsymbol{y}_{obs}|\boldsymbol{\Delta})$, implying that the parameter of interest, $\boldsymbol{\Delta}$, can be inferred directly from the observed data.

Rubin (1976) provides sufficient conditions under which ignoring the missing data mechanism still yields correct direct likelihood inference about $\boldsymbol{\Delta}$. The conditions are: (1) Satisfying missing at random assumption (MAR), i.e., for each value of φ, $h_{\varphi}(\boldsymbol{m}|\boldsymbol{y}_{obs}, \boldsymbol{y}_{mis}) = h_{\varphi}(\boldsymbol{m}|\boldsymbol{y}_{obs})$ for all values of $\boldsymbol{y}_{mis}$; (2) The parameter φ is distinct from $\boldsymbol{\Delta}$, which means that all possible values of φ are possible in combination with all possible values of $\boldsymbol{\Delta}$.

In what follows, we will discuss the missing mechanisms induced by the two routing rules using Rubin's (1976) framework. One is based on $\hat{\theta}$ which is used in the current operational testing, and the other is based on true θ which is certainly unrealistic. The rationale for considering the latter design is that several previous studies used multiple-group calibration approach to estimate item parameters from the MST design but they were unsuccessful (e.g., Lu et al., 2017; Cai et al., 2018). Therefore, we intend to provide a theory grounded argument that only when the routing is based on true θ that the multiple-group approach is needed. This argument is also further backed up by the simulation results in section 5.

4.1 Routing based on $\hat{\theta}$

We first consider a MST design where the routing rule is based on interim $\hat{\theta}$, which is estimated from the responses and the previously known item parameters[3] in the routing block. Under this design, the marginal likelihood of $\boldsymbol{\Delta}$ for person *i* by integrating out both $\boldsymbol{y}_{i,mis}$ and θ is

[3] These are the initially estimated item parameters obtained from the previous administrations. The item parameters will be recalibrated again with the MST-generated data, which is the routine analysis in NAEP to avoid any aberrances due to item parameter drift.

$$\int_{\theta} \int_{\boldsymbol{y}_{i,mis}} L_{\varphi}\left(\theta, \boldsymbol{\Delta}|\boldsymbol{y}_{i,obs}, \boldsymbol{y}_{i,mis}, \boldsymbol{m}_i\right) d\boldsymbol{y}_{i,mis} d\theta$$

$$= \int_{\theta} \int_{\boldsymbol{y}_{i,mis}} L\left(\theta, \boldsymbol{\Delta}|\boldsymbol{y}_{i,obs}, \boldsymbol{y}_{i,mis}\right) h_{\varphi}\left(\boldsymbol{m}_i|\boldsymbol{y}_{i,obs}, \boldsymbol{y}_{i,mis}, \theta\right) d\boldsymbol{y}_{i,mis} d\theta$$

$$= \int_{\theta} \int_{\boldsymbol{y}_{i,mis}} L\left(\theta, \boldsymbol{\Delta}|\boldsymbol{y}_{i,obs}, \boldsymbol{y}_{i,mis}\right) h_{\varphi}\left(\boldsymbol{m}_i|\boldsymbol{y}_{i,obs}^{R}\right) d\boldsymbol{y}_{i,mis} d\theta, \quad (9)$$

where $\boldsymbol{y}_{i,obs}^{R}$ denotes the observed responses on the items in the routing block. Here, φ contains the pre-specified cut-offs for routing decisions and therefore it is distinct from the target parameters $\boldsymbol{\Delta}$. Because the missing data mechanism only depends on observed data, the MAR assumption is automatically satisfied. Then, the last equality in Eq. (9) holds, and $h_{\varphi}(\boldsymbol{m}_i|\boldsymbol{y}_{i,obs}^{R})$ indicates the missing data process that depends on the observed response vector because $\hat{\theta}$ is estimated from $\boldsymbol{y}_{i,obs}^{R}$.

Further expand $L\ (\theta, \boldsymbol{\Delta}|\boldsymbol{y}_{i,obs}, \boldsymbol{y}_{i,mis}) = \left[\prod_j P\left(\boldsymbol{y}_{i,obs}|\theta, \Delta_j\right)\right] g(\theta|\mu_\theta = 0, \sigma_\theta^2 = 1) P(\boldsymbol{y}_{i,mis}|\theta)$ in Eq. (9) such that it can be simplified as

$$\int_{\theta} \left[\prod_j P\left(\boldsymbol{y}_{i,obs}|\theta, \Delta_j\right)\right] g(\theta|\mu_\theta = 0, \sigma_\theta^2 = 1) h_{\varphi}(\boldsymbol{m}_i|\boldsymbol{y}_{i,obs}^{R}) d\theta. \quad (10)$$

This is because $\int_{\boldsymbol{y}_{i,mis}} P(\boldsymbol{y}_{i,mis}|\theta)\ d\boldsymbol{y}_{i,mis} = 1$. So the MML item calibration method intends to maximize the marginal likelihood

$$\prod_{i=1}^{N} \int_{\theta} \left[\prod_j P\left(\boldsymbol{y}_{i,obs}|\theta, \Delta_j\right)\right] g(\theta|\mu_\theta = 0, \sigma_\theta^2 = 1) h_{\varphi}(\boldsymbol{m}_i|\boldsymbol{y}_{i,obs}^{R}) d\theta$$

$$= \prod_{i=1}^{N} h_{\varphi}(\boldsymbol{m}_i|\boldsymbol{y}_{i,obs}^{R}) \int_{\theta} \left[\prod_j P\left(\boldsymbol{y}_{i,obs}|\theta, \Delta_j\right)\right] g\left(\theta|\mu_\theta = 0, \sigma_\theta^2 = 1\right) d\theta. \quad (11)$$

When both the MAR assumption and distinctiveness assumption are satisfied, Rubin (1976)'s ignorability condition is satisfied. Hence, a single-group marginal maximum likelihood (MML)

introduced in section 3.1 is sufficient for item calibration in this MST design. Indeed, after taking a log-transformation of Eq. (11), the term $h_{\varphi}(\boldsymbol{m_i}|\boldsymbol{y}_{i,obs}^{r})$ is no longer relevant because it does not contain $\boldsymbol{\Delta}$. In this case, maximizing Eq. (11) is equivalent to maximizing Eq. (3).

We can also show that for the EM algorithm in section 3.2, when the MAR assumption is satisfied, the EM algorithm can proceed based solely on the observed data. The detailed derivation is provided in the Appendix.

4.2 Routing based on true θ

Several recent studies (Cai, et al., 2018; Lu et al., 2018) have used the multiple-group MML method for MST item calibration and found biased parameter estimates. In this section, we will show that, from missing data principle, the multiple-group calibration approach is only appropriate for a special, unrealistic, scenario where the routing is based on true θ. Even so, the group specific θ distribution also needs to be defined correctly.

In the current practice, multiple-group MML proceeds by assuming θ distribution follows normal $N(\mu_g, \sigma_g^2)$, where g denotes the gth group (Cai, et al., 2011). There are two commonly used approaches to remove the scale indeterminacy. The first approach is to let the mean and variance for all three groups be estimable parameters with the constraints that the overall mean and standard deviation are 0 and 1 respectively (Lu, Jia, & Wu, 2017). The second approach is to fix the mean and variance of θ in one group to constants, and let them in all remaining groups to be freely estimated.

According to the discussion in 4.1, when routing is based on $\hat{\theta}$, the ignorablility condition is satisfied and hence a single-group MML is sufficient. Using multiple-group MML not only adds estimation complexity due to additional parameters, but it is also based on a false

assumption that θ distribution for each subgroup follows a normal distribution. This is exactly the reason why the previous studies using the multiple-group calibration were unsuccessful. There is one exception when the multiple-group MML is necessary. That is when the routing decision is made based on true θ. For example, let c_1 and c_2 be the two cut-offs along the θ continuum, and now the missing mechanism is

$$h_{\varphi}(\boldsymbol{m}|\theta) = \begin{cases} 1_{\theta>c_2} & \text{if this person takes the difficult block} \\ 1_{c_1\le\theta\le c_2} & \text{if this person takes the medium block} \\ 1_{\theta<c_1} & \text{if this person takes the easy block} \end{cases} . \quad (12)$$

The MAR assumption is no longer satisfied because the missing data depends on the unknown latent variable θ which itself is also missing. Replacing $h_{\varphi}\left(\boldsymbol{m}_i\middle|\boldsymbol{y}_{i,obs}^{R}\right)$ in Eq. (10) by (12) results in a marginal likelihood that is comprised of three components,

$$\prod_{i\in\text{difficult}}\int_{\theta}\left[\prod_j P\left(\boldsymbol{y}_{i,obs}|\theta,\Delta_j\right)\right]g\left(\theta\middle|\mu_\theta=0,\sigma_\theta^2=1\right)1_{\theta>c_2}d\theta\ \times$$

$$\prod_{i\in\text{medium}}\int_{\theta}\left[\prod_j P\left(\boldsymbol{y}_{i,obs}|\theta,\Delta_j\right)\right]g\left(\theta\middle|\mu_\theta=0,\sigma_\theta^2=1\right)1_{c_1\le\theta\le c_2}d\theta\ \times$$

$$\prod_{i\in\text{easy}}\int_{\theta}\left[\prod_j P\left(\boldsymbol{y}_{i,obs}|\theta,\Delta_j\right)\right]g\left(\theta\middle|\mu_\theta=0,\sigma_\theta^2=1\right)1_{\theta<c_1}d\theta \quad (13)$$

It is clear from Eq. (13) that a three-group calibration needs to be performed, and each group has a θ distribution that follows a truncation of a standard normal distribution.

4.3 The challenge of calibration by subscale

Many large scale assessments such as NAEP or PISA (e.g., Liu, Wilson, & Paek, 2008) measure students' performance on multiple subscales within a given subject domain. The standard practice of NAEP item calibration is to calibrate items from each subscale separately using the traditional single-group MML method (Wu & Lu, 2017; Wu & Xi, 2017). However, this procedure yields biased item parameter estimates when the response data are collected from

the MST design (e.g., Lu et al., 2017). Previous studies have neither given a justifiable explanation nor provided a viable solution.

In fact, from the missing data theory, it can be easily verified that when the calibration is conducted per subscale, the MAR assumption is violated. This is because, by design, the missing data mechanism is based on the observed responses from all items in the routing block, i.e., $h_{\varphi}\left(\boldsymbol{m}_i|\boldsymbol{y}_{i,obs}^{R}\right)$. However, if one conducts the calibration per subscale, for subscale *d*, we have the marginal likelihood for person *i* as follows,

$$\int_{\theta^d}\int_{\boldsymbol{y}_{i,mis}^{d}} L_{\varphi}\left(\boldsymbol{\Delta},\theta^{d}|\boldsymbol{y}_{i,obs}^{d},\boldsymbol{y}_{i,mis}^{d},\boldsymbol{m}_i\right)d\boldsymbol{y}_{i,mis}^{d}d\theta^{d}$$

$$=\int_{\theta^d}\int_{\boldsymbol{y}_{i,mis}^{d}} L\left(\boldsymbol{\Delta},\theta^{d}|\boldsymbol{y}_{i,obs}^{d},\boldsymbol{y}_{i,mis}^{d}\right)h_{\varphi}\left(\boldsymbol{m}_i|\boldsymbol{y}_{i,obs}^{R,d}\right)d\boldsymbol{y}_{i,mis}^{d}d\theta^{d}. \quad (14)$$

In Eq. (14), $\boldsymbol{y}_{i,obs}^{R,d}$ denotes the observed responses from person *i* on items in subscale *d* in the routing block. Please note that because the missing data function $h_{\varphi}(\boldsymbol{m}_i|\boldsymbol{y}_{i,obs}^{R,d}) \neq h_{\varphi}\left(\boldsymbol{m}_i|\boldsymbol{y}_{i,obs}^{R}\right)$, using (14) will inevitably introduce bias due to the misspecification of the missing data function. Indeed, if let $\boldsymbol{y}_{i,obs}^{R} \equiv \left(\boldsymbol{y}_{i,obs}^{R,d},\boldsymbol{y}_{i,obs}^{R,-d}\right)$, where $\boldsymbol{y}_{i,obs}^{R,-d}$ denotes the observed responses from person *i* on all items in the routing block except subscale *d* , then if one performs the calibration by subscale via MML following Eq. (14), $\boldsymbol{y}_{i,obs}^{r,-d}$ is considered as "missing" data because it is not used in the calibration. Therefore, the missing data actually depends on the "missing" observations, violating the missing at random assumption. Following this argument, a simple solution is to augment the subscale data $\boldsymbol{y}_{i,obs}^{d}$ by $\boldsymbol{y}_{i,obs}^{R,-d}$, and the MAR assumption will be satisfied such that a single-group MML still applies.

Figure 2 provides an illustrative comparison of the traditional calibration per subscale approach, and our proposed, modified approach. Assuming the test contains three subscales, for

the modified approach, although item responses from the other two subscales in the routing block are also used in item calibration, the item parameters for those subscales are considered "nuisance".

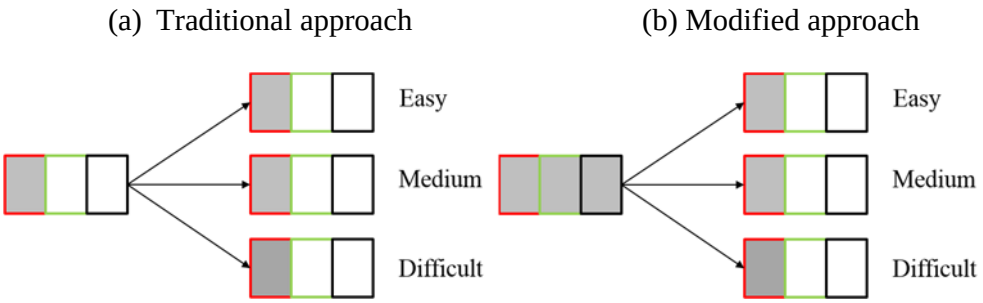

Figure 2. Illustration of calibration per subscale. The three boxes with different colored lines represent three different subscales. If one intends to calibrate item parameters from scale 1 (red color), item responses from the shaded area are used as input.

5. Simulation Studies

Two simulation studies were conducted to evaluate the performance of the different calibration methods under a typical NAEP design. The 2PL model was used throughout the simulation studies because its item parameters tend to be relatively easy to recover, whereas the *c*-parameter estimation in the 3PL model is known to be challenging (Thissen & Wainer, 1982; Swaminathan & Gifford, 1986).

5.1 Design and Methods

Item bank and MST design The items were obtained from NAEP 2011 Grade 8 mathematics assessment[4]. The item bank was constructed by pooling together items in all five content areas and all testing blocks. There were 115 items in total, from which four testing modules were assembled. For content balancing purpose, the following procedure was conducted

[4] The real item parameters were retried from https://nces.ed.gov/nationsreportcard/tdw/analysis/scaling_irt_math.aspx

on each of the five subscales. First, the items were ranked in ascending order in terms of the discrimination parameter. About ¼ of the items with the lowest a values were chosen to form the routing module. Selecting items with lower-*a* parameters at the beginning of the test is consistent with the suggestions in Chang and Ying (1996). This design not only helps balance item usage and but also makes the test more robust to random errors (e.g., incorrect answers due to test anxiety) at the beginning of the test (Chang & Ying, 2008). Then, the remaining items were ordered by the difficulty parameter, and an "easy" module is made from 1/3 of the remaining items with the lowest difficulty. Similarly, the "difficult" module consists of 1/3 of the most difficult items. The final module consists of the last 1/3 items with medium difficulty. Table 1 shows the number of items in each module and each subscale. Although the number of items per subscale differs by test design, they are roughly evenly distributed across four modules. Table 2 presents the descriptive statistics of the item parameters.

Table 1. Item distributions per module and per subscale

	Number sense, properties, and operations	Measurement	Geometry and spatial sense	Data analysis, statistics, and probability	Algebra and functions	Total
Routing	5	6	7	4	9	31
Easy	5	5	6	4	9	29
Medium	4	5	6	3	9	27
Hard	5	5	6	3	9	28
Total	19	21	25	14	36	115

Table 2. Descriptive statistics of item parameters

	Mean		SD	
	a	*b*	*a*	*b*
Routing	0.63	-0.01	0.14	1.24
Easy	1.05	-0.34	0.24	0.41
Medium	1.08	0.46	0.27	0.38
Hard	1.21	1.24	0.37	0.48
Total	0.98	0.32	0.34	0.94

Response generation and routing Two simulation designs (denoted as Design I and Design II) were considered depending upon the routing methods (routing based on true θ vs. routing based on $\hat{\theta}$). Sample size was set at 3,000. In both designs, every simulee responded to the items in the routing module, and roughly 1/3 of the simulees were routed to one of the three target modules based on the routing rules. The 1/3 and 2/3 quantiles of the standard normal distribution were chosen as the two fixed cut points, and they are $c_1 = -.43$ and $c_2 = .43$.

For design I, a group of 3000 simulees' true θs was generated from a standard normal distribution. Then the responses were generated based on 2PL in Eq. (1). The next module was decided by the location of the simulees' **true** θs relative to the cut points. If their true θs were smaller than c_1, they were assigned to the easy module; if their true θs were larger than c_2, they were assigned to the difficult module; and if their true θs were between the two cut points, they were assigned to the medium module. This design, although unrealistic in practice, results in a missing not at random (MNAR) condition.

Design II only differs from Design I by the routing method. To reduce random error, Design II shared the same 3000 θs and the same responses from the routing block in Design I. After the routing stage, individual θ was estimated via the expected a posterior (EAP) with a standard normal prior[5], and the next module was decided by the location of $\hat{\theta}$ relative to the cut points. This design results in a MAR condition.

Calibration methods Table 3 summarizes the calibration methods used in the two simulation designs. If viewing all items in the test measure a single, unidimensional trait, five

[5] See Eq (2.1) on page 7 of the following document https://www.nagb.gov/content/nagb/assets/documents/publications/achievement/developing-achievement-levels-2011-naep-grade8-grade12-writing-technical-report.pdf

different methods were compared. They are (1) the single-group MML (denoted as S-MML hereafter) assuming the entire calibration sample as a single group with θ from a standard normal distribution; (2) the multiple-group MML with all normal (denoted as M-MML-N), where we assume the population consists of three subpopulations, all of which follow a normal distribution with group specific mean and variance. Here the mean and variance for the middle group were fixed at their true values to fix the scale. We considered this method jus to replicate the studies by Cai et al. (2018) and Lu et al. (2018); (3) the multiple-group MML with truncated normal (denoted as M-MML-T) according to the description in section 4.2; (4) the single-group fixed parameter calibration (S-FPC) and (5) the multiple-group FPC (M-FPC). With FPC, the calibration proceeds in two steps. In the first step, the complete response matrix from the routing block were calibrated as usual, then those routing item parameters were fixed at their estimated values, and the targeted block items were calibrated via FPC. By single-group, we refer to estimating π_k's as if they are from a single population, whereas by multiple-group, we refer to estimating π_k's separately for three subpopulations. It is anticipated that when the MAR assumption is satisfied with $\hat{\theta}$ routing, all single-group methods should outperform the multiple-group methods. On the other hand, when the MAR assumption is violated with true θ routing, the multiple-group methods should be preferred.

With respect to calibration per subscale, again both single-group and multiple-group approaches were evaluated. Within the single-group framework, we considered both the MML and FPC methods, and for each method, we considered two scenarios: the one with all routing items (i.e., our modified approach that satisfies the MAR assumption) and the one with the routing items only pertinent to the corresponding subscales (i.e., current method). These 2 (MML vs. FPC) by 2 (All vs. Only) result in four methods, denoted as S-MML-All, S-FPC-All, S-

MML-Only, and S-FPC-Only hereafter. Regarding the multiple group approach, we only considered the FPC method because it does not need to specify the distribution of θ in advance. They are referred to as multiple-group FPC with only subscale relevant routing items (M-FPC-Only) and with all routing items (M-FPC-All).

30 replications were conducted per condition, and two prior distributions of the item parameters (i.e., $\log(a) \sim N(0, 0.5^2)$ and $b \sim N(0, 2^2)$) under the 2PL model were specified for effective runs of the FPC method. These are the default priors used in BILOG-MG and PARSCALE (Kim, 2006, p. 357).

Table 3. Summary of the calibration methods for different simulation designs[6]

Single vs. Multiple group	Methods	Notation	Simulation I Unidimensional 2PL with θ routing	Simulation II Unidimensional 2PL with $\hat{\theta}$ routing
Scenario 1: All items are calibrated on a single scale				
S	MML	S-MML	√	√
M	MML with all normal	M-MML-N	√	√
M	MML with truncated normal	M-MML-T	√	√
S	Fixed parameter EM (FPC)	S-FPC	√	√
M	Fixed parameter EM (FPC)	M-FPC	√	√
Scenario 2: Items from different content areas (i.e., subscales) are calibrated on separate scales				
S	MML per subscale	S-MML-Only		√
S	Modified MML per subscale	S-MML-All		√
S	Fixed parameter EM per subscale	S-FPC-Only		√
S	Modified Fixed parameter EM per subscale	S-FPC-All		√
M	Fixed parameter EM per subscale	M-FPC-Only		√
M	Modified Fixed parameter EM per subscale	M-FPC-All		√

[6] The R and MATLAB source code for running all the proposed methods can be found on https://sites.uw.edu/pmetrics/publications-and-source-code/

5.2 Results

Overall unidimensional calibration The evaluation criteria are the average bias and root mean squared error (RMSE) of the *a*- and *b*- parameters. They were computed first across all replications per item, and then averaged over all items. The parameter recovery were summarized for both all items and items within each block. Table 4 presents the average bias and RMSE for design I with true θ routing and Table 5 presents the item parameter recovery for design II with estimated $\hat{\theta}$ routing.

Table 4. Average bias and RMSE of *a*- and *b*- parameters with 2PL model calibration for Design I (i.e., true θ routing)

	All		Routing		Easy		Medium		Hard	
Method	*a*	*b*	*a*	*b*	*a*	*b*	*a*	*b*	*a*	*b*
bias										
S-MML	-0.28	0.25	0.01	-0.03	-0.30	0.42	-0.61	0.78	-0.247	-0.13
M-MML-N	0.00	0.01	0.00	0.02	-0.06	0.04	0.04	0.00	0.02	-0.01
M-MML-T	0.01	0.02	0.04	0.00	0.00	0.08	-0.03	0.05	0.01	-0.08
S-FPC	-0.27	0.21	0.00	0.01	-0.30	0.45	-0.55	0.49	-0.28	-0.09
M-FPC	-0.05	0.03	0.00	0.01	-0.05	0.05	-0.10	0.07	-0.05	0.01
RMSE										
S-MML	0.30	0.55	0.04	0.08	0.31	0.45	0.62	1.58	0.26	0.17
M-MML-N	0.12	0.08	0.03	0.05	0.14	0.10	0.20	0.11	0.11	0.05
M-MML-T	0.14	0.12	0.05	0.08	0.16	0.14	0.24	0.20	0.12	0.09
S-FPC	0.29	0.31	0.04	0.07	0.32	0.47	0.56	0.59	0.29	0.15
M-FPC	0.13	0.10	0.04	0.07	0.15	0.12	0.22	0.15	0.12	0.06

Table 5. Average bias and RMSE of *a*- and *b*- parameters with 2PL model calibration for Design II (i.e., estimated $\hat{\theta}$ routing)

	All		Routing		Easy		Medium		Hard	
Method	*a*	*b*	*a*	*b*	*a*	*b*	*a*	*b*	*a*	*b*
bias										
S-MML	0.01	-0.03	0.01	-0.03	0.01	-0.03	0.00	-0.03	0.00	-0.04
M-MML-N	-0.10	0.01	-0.26	0.02	-0.10	-0.85	0.07	-0.03	-0.10	0.94
M-MML-T	0.35	-0.10	0.06	0.00	0.35	-0.18	0.79	-0.20	0.26	-0.03
S-FPC	-0.02	0.02	0.00	0.01	-0.02	0.03	-0.02	0.02	-0.04	0.01
M-FPC	0.41	-0.10	0.00	0.01	0.39	-0.32	1.01	-0.21	0.31	0.10
RMSE										
S-MML	0.10	0.08	0.03	0.08	0.12	0.09	0.14	0.09	0.11	0.07
M-MML-N	0.16	0.63	0.26	0.64	0.13	0.85	0.14	0.07	0.13	0.94
M-MML-T	0.38	0.15	0.07	0.09	0.39	0.20	0.83	0.23	0.29	0.08
S-FPC	0.10	0.08	0.04	0.07	0.11	0.09	0.13	0.08	0.11	0.06
M-FPC	0.44	0.19	0.04	0.07	0.43	0.33	1.04	0.25	0.34	0.13

Several conclusions can be drawn from Tables 4 and 5. First and unsurprisingly, the MML and FPC methods, including both of their single-group and multiple-group versions, perform similarly in all conditions. That is, when S-MML performs well, S-FPC also performs well. In contrast, when M-MML (both M-MML-N and M-MML-T) performs badly, so does M-FPC. This indicates that one can either concurrently calibrate all items or calibrate routing items first and targeted block items second. Second and more interestingly, when routing is based on true θ, multiple-group approach outperforms single-group approach regardless of the specific calibration method. In this case, the MAR assumption is violated, and using a single-group approach based on observed data ignores the missing data mechanism. As a result, the item parameter estimates are severely biased. On the other hand, when routing is based on estimated $\hat{\theta}$, then the single-group approach performs much better than the multiple-group approach. In the latter case, the items in the routing block are still recovered well, it is the targeted blocks that are adversely affected.

Calibration per subscale Although the items are calibrated separately per subscale, the same evaluation criteria were still used to summarize the parameter recovery. In this case, only simulation design II was considered because they mimic the real practice closely. Tables 6 reports the results for Design II.

Table 6. Average bias and RMSE of *a*- and *b*- parameters with 2PL model calibration per subscale for Design II (i.e., estimated $\hat{\theta}$ routing)

	All		Routing		Easy		Medium		Hard	
Method	*a*	*b*	*a*	*b*	*a*	*b*	*a*	*b*	*a*	*b*
bias										
S-MML-Only	-0.35	0.25	-0.04	0.00	-0.39	0.74	-0.67	0.63	-0.34	-0.33
S-MML-All	0.01	0.00	0.00	0.00	0.02	0.00	0.01	0.00	0.01	-0.01
S-FPC-Only	-0.29	0.19	0.00	0.01	-0.34	0.71	-0.54	0.40	-0.30	-0.35
S-FPC-All	-0.01	0.02	0.00	0.01	-0.01	0.02	-0.02	0.02	-0.03	0.01
M-FPC-Only	0.20	-0.08	0.00	0.01	0.24	-0.27	0.44	-0.16	0.16	0.11
M-FPC-All	0.73	-0.13	0.00	0.01	0.70	-0.38	1.95	-0.27	0.39	0.09

RMSE										
S-MML-Only	0.36	055	0.06	0.10	0.40	0.76	0.67	1.00	0.36	0.37
S-MML-All	0.12	0.08	0.04	0.07	0.15	0.09	0.18	0.10	0.12	0.06
S-FPC-Only	0.31	0.42	0.04	0.07	0.35	0.72	0.55	0.53	0.32	0.39
S-FPC-All	0.11	0.08	0.04	0.07	0.13	0.09	0.16	0.09	0.13	0.07
M-FPC-Only	0.29	0.17	0.04	0.07	0.34	0.29	0.55	0.19	0.25	0.14
M-FPC-All	0.79	0.23	0.04	0.07	0.77	0.39	2.05	0.33	0.43	0.13

It is shown from Table 6 that, consistent with prior findings (e.g., Lu et al., 2017), using a single-group MML or a single-group FPC per subscale calibration leads to severe bias. This is due to the violation of the MAR assumption. The modified approach, however, by augmenting the subscale item responses by responses on all routing items, help satisfy the MAR assumption. Therefore, as expected, the modified approach greatly improves estimation accuracy. Both S-MML-All and S-FPC-All result in almost unbiased parameter estimates. Another interesting finding worth mentioning is, when the MAR assumption is violated, the multiple group approach outperforms the single group approach. This is reflected in the better results from M-FPC-Only than from S-FPC-Only, although M-FPC-Only still yields large bias and RMSE relatively. One explanation is, the number of items per subscale per block (see Table 1) is too few to help recover the underlying θ distribution per group in the M-FPC-Only approach. Further simulation studies need to be conducted to verify the conjecture.

6. Real Data Analysis

The real response data from a special NAEP MST grade 8 math assessment study in year 2011 is used as an example. The total sample size is 8,401, in which about 40% of the students (N_2 = 3344) were placed in the *experiment* sample (taking the two-stage MST, see Figure 1), and roughly 60% (N_1 = 5057) were in the *calibration* sample (random routing). In the routing stage, there are two parallel forms, and examinees were randomly assigned to one of the two forms,

hence the missing data in the routing stage is completely at random. Table 7 presents the sample size per form and per target block from each sample. As shown, the sample sizes are comparable across different forms/blocks, and the sample size is enough to calibrate the 2PL model parameters accurately. Table 8 presents the number of items per content domain within each form/block.

Table 7. Sample size per form/block

Routing Form	1			2			
Target block	Easy	Medium	Hard	Easy	Medium	Hard	Total
Experiment sample	669	715	273	681	734	272	3344
Calibration sample	847	826	868	857	848	811	5057
Total	1516	1541	1141	1538	1582	1083	8401

Table 8. Number of items per content domain in each form/block.

	Routing		Target		
	Form 1	Form 2	Easy	Medium	Hard
Number properties and operations	3	4	2	2	3
Measurement	3	3	2	2	2
Geometry	3	3	3	3	3
Data analysis statistics and probability	2	2	2	1	2
Algebra	6	5	5	5	3

For both samples, two scenarios were considered, i.e., items from the entire test were calibrated on a single scale (labeled as “overall calibration” in Table 9) and items from each content area were calibrated on separate scales (labeled as “calibration per subscale”). For overall calibration, four methods are compared. They are the single group maximum likelihood estimation (S-MML), single-group EM (S-EM), single-group fixed parameter calibration (S-FPC), and multiple group FPC (M-FPC). The multiple-group MML is not considered because the FPC method is more flexible to model the different shapes of θ distributions per group. It is expected that all four approaches will produce similar item parameter estimates when data comes

from the calibration sample, whereas M-FPC will produce biased item parameter estimates when the data comes from experiment sample.

For the calibration per subscale, which is more interesting, two approaches are compared as shown in Table 9. They are both single-group methods because the multiple-group alternatives did not produce satisfactory results according to the simulation findings. Again, both methods should work reasonably well on the calibration sample, whereas only the S-FPC-All method is expected to produce comparable and almost unbiased item parameters using the experiment sample. The S-MML-Only and the S-MML-All methods evaluated in the simulation study are no longer considered here for two reasons: (1) Both of them perform similarly to the FPC alternatives when the population distribution of θ is normal; and (2) the distribution of θ in the current sample departs slightly from normal (see Figures 3 and 4) and hence FPC is preferred.

Table 9. Calibration plan for the real data

	Overall calibration				Calibration per subscale	
Calibration/Experiment sample	S-MML	S-EM	S-FPC	M-FPC	S-FPC-Only	S-FPC-All

6.1 Overall calibration results

Figures 3 and 4 present the scatter plots of the estimated item *a*- and *b*- parameters from pairs of methods for the two samples respectively. In both Figures, the single-group EM method serves as the benchmark method because as discussed earlier, the missing data in this scenario could be considered MAR. As shown in Figure 3, the item parameter estimates from all four methods align well when the data is from the calibration sample. Note that S-EM, S-FPC, and M-FPC allow flexible (non-parametric) θ distributions whereas S-MML implicitly assumes a normal θ distribution. There is a slight misalignment between the estimated a-parameters from S-MML versus the estimates from S-EM, which implies that the θ distribution in the calibration sample does not strictly follow a normal distribution. This misalignment is exacerbated the

experiment sample. Moreover, in Figure 4, both S-EM and S-FPC produce similar parameter estimates, whereas the item parameter estimates from M-FPC do not align well. This is consistent with the simulation findings. In addition, because the θ distribution in the calibration and experiment samples do not seem to be the same, the item parameter estimates from these two samples may not be directly comparable, resulting in a slight misalignment in Figure 5. Given this observation, the comparison between the two samples will be dropped from further discussion.

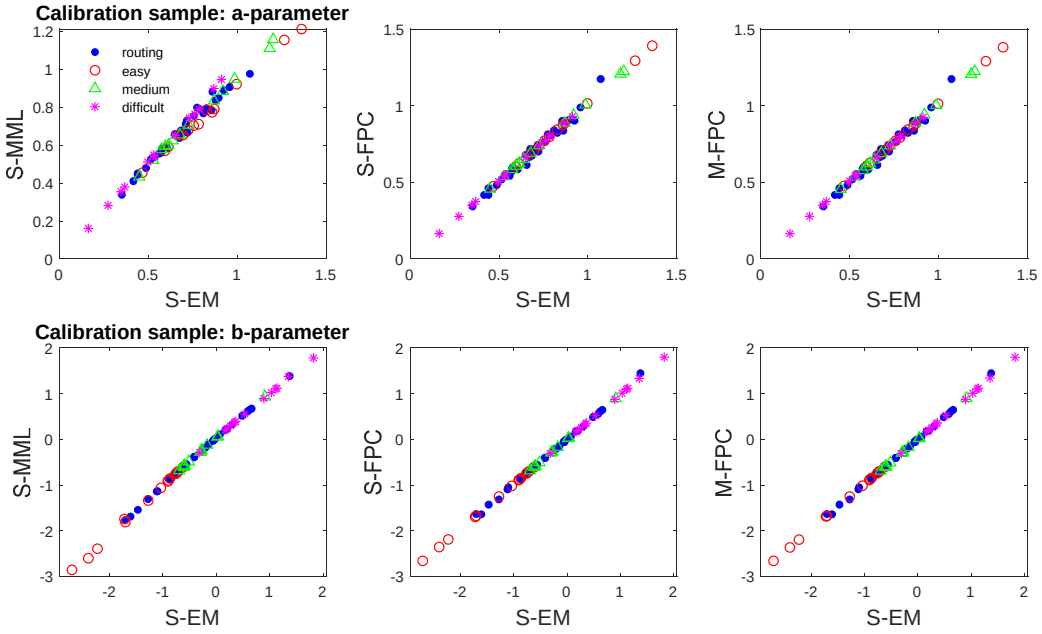

Figure 3. Scatter plots of the estimated item *a*- and *b*- parameters from overall calibration for the calibration sample

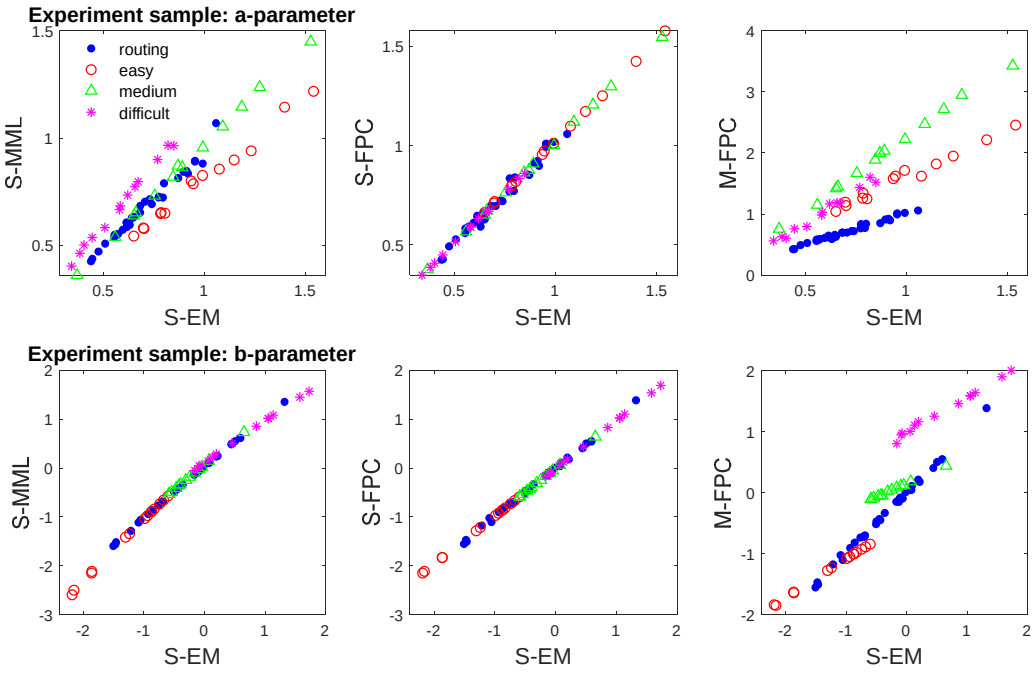

Figure 4. Scatter plots of the estimated item *a*- and *b*- parameters from overall calibration for the experiment sample

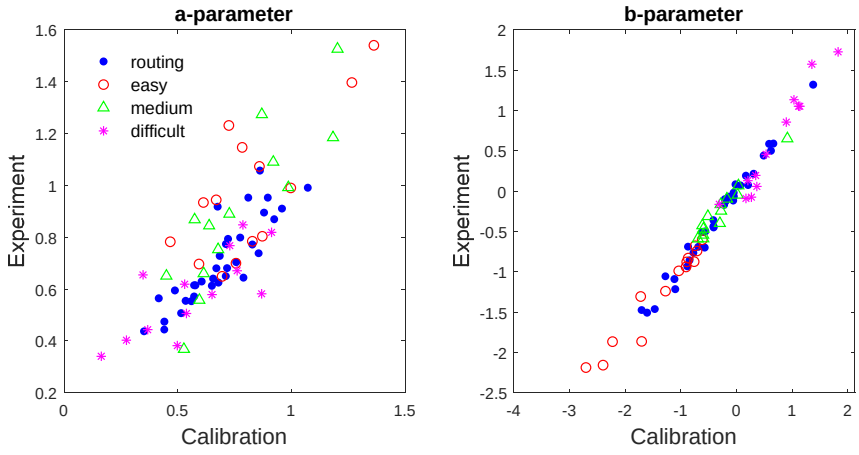

Figure 5. Scatter plots of the estimated item *a*- and *b*- parameters from calibration vs. experiment sample using the S-EM algorithm

6.2 Calibration per subscale results

This section includes the results of calibrating items from different content areas on their respective scales. Figure 6 presents the results for the calibration sample, comparing the S-FPC-Only and S-FPC-All methods against S-EM which is again the benchmark. Similar findings emerge. That is, the S-FPC-All method generates item parameter estimates that are in closer alignment with the S-EM approach, whereas the S-FPC-Only approach produces biased item parameter estimates. The biases are much more extreme when evaluating the results from the experiment sample, as reflected in Figure 7. This observation further reinforces that our proposed S-FPC-All approach should be preferred to the original S-FPC-Only approach because it reinstates the MAR assumption.

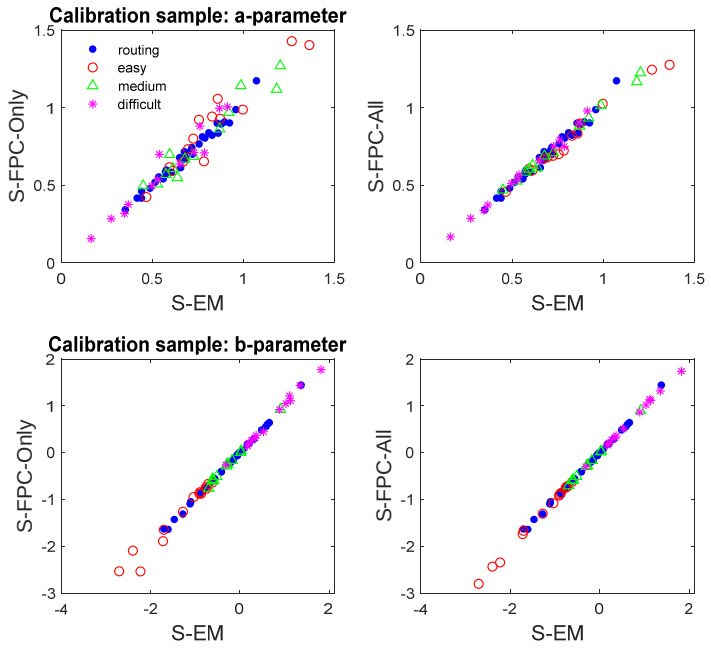

Figure 6. Scatter plots of the estimated unidimensional item *a*- and *b*- parameters from the calibration sample

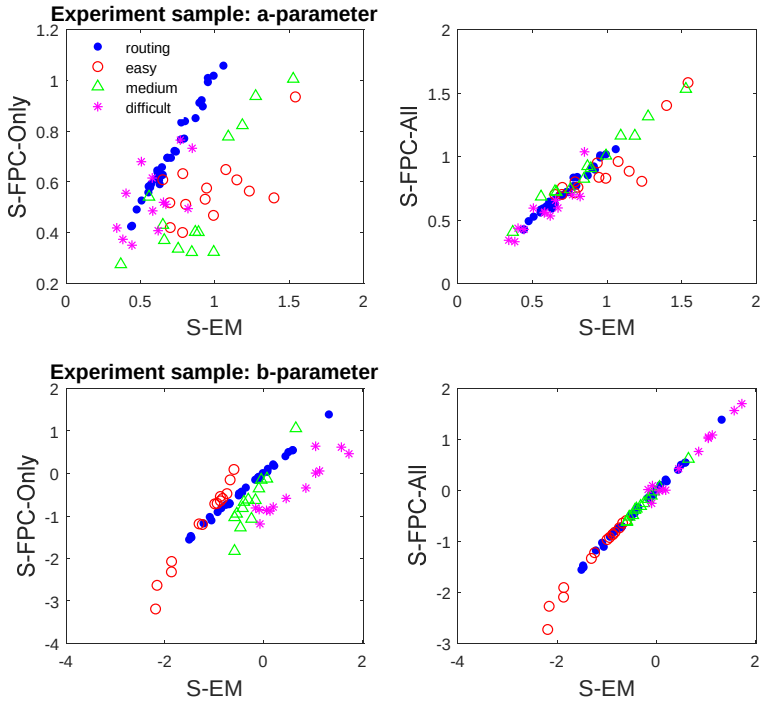

Figure 7. Scatter plots of the estimated unidimensional item *a*- and *b*- parameters from the experiment sample

7. Discussion

Multistage testing design has recently emerged as a powerful test delivery mode because it can help measure the high-achieving and low-achieving subgroups more accurately than the traditional linear forms (Yan, von Davier, & Lewis, 2014). On the other hand, compared to computerized adaptive testing that is fully adaptive at item level, MST contains pre-assembled forms such that the various constraints in the test blueprint can be checked in advance. In practice, if the same items are used over a long period of time, the parameters of those items are often recalibrated to check potential parameter drift. Item calibration is an important step in any IRT based scoring and inference. Any biases introduced in item calibration will propagate in subsequent steps and consequently bias the conclusions that may have profound policy

relevance. Only when the item parameters are precisely calibrated and linked across years can long-term trend lines be constructed and subgroup comparisons made.

Questions remain as to how to calibrate items using the incomplete data from the MST design. Complication arises when there are multiple correlated subscales per assessment, and when it is necessary to put item parameters on their respective subscale score reporting metric. Although several recent studies have started to explore various item calibration methods with the MST design (e.g., Lu et al., 2017, 2018, Cai, et al., 2018, Jewsbury & van Rijn, 2018), they have not thoroughly analyzed the MST calibration challenge from a missing data perspective. For example, Lu et al. (2018) tried to provide different priors on *a*- and *b*- parameters to bring down the estimation bias, but there was not much success. Therefore, it remains unclear why the multiple-group EM does not produce an acceptable parameter recovery. In addition, a viable method is needed to properly calibrate item parameters per subscale.

In this paper, we draw upon Rubin (1976)'s missing data theory, and explicitly show that when the routing decision is based on $\hat{\theta}$, the ignorability condition (i.e., MAR and distinctiveness assumption) is satisfied such that the as-usual, single-group calibration methods are sufficient. Using a multiple-group approach, however, will introduce additional bias regardless of the actual calibration methods. On the other hand, when the MAR assumption is violated, as in the true θ routing condition, the multiple group approach is necessary. As an additional check, Table 10 presents the "misclassification" rate from the simulation design II. The true group membership is based on comparing an individual's true θ to the two cut-offs, whereas the assigned group membership is based on estimated $\hat{\theta}$. Although there is only about 20% discrepancy on average, the same calibration method can perform drastically different in the two scenarios (true θ vs. estimated $\hat{\theta}$ routing), as reflected by results in Tables 4 & 5. This

reinforces the importance of checking the MAR assumption. In fact, prior studies (e.g., Mislevy & Wu, 1996; Glas, 2010; Eggen & Verhelst, 2011) have concluded that the MAR assumption is satisfied for MST design when the focus is on θ-estimation given known item parameters (Mislevy & Sheena, 1989), or on item calibration (Glas, 2010; Eggen & Verhelst, 2011). Following this perspective, we propose a simple, yet effective method to resolve the calibration by subscale challenge. The key is to augment the response data such that MAR assumption is satisfied.

Table 10. Misclassification rate from the simulation design II

True Group based on θ	Assigned Group based on $\hat{\theta}$		
	Easy	Medium	Difficult
Easy	.830	.166	.002
Medium	.140	.725	.135
Difficult	0	.157	.845

In this paper, three mainstream calibration methods are reviewed and discussed in the context of missing data, they are MML, EM, and FPC. While MML often assumes θ follows a normal distribution or other known parametric distributions, the EM algorithm can naturally handle the case when the parametric form of the θ distribution is unknown. This is because it directly estimates the probability mass function of θ by treating it as a discrete random variable. This feature is extremely useful in particular within the FPC framework because when certain item parameters are fixed, the entire θ distribution can be freely estimated. For instance, in the simulation design I when routing is based on true θ, both the multiple-group MML with normal (M-MML-N) and multiple-group MML with truncated normal (M-MML-T) methods assume the shape of θ distribution per group is known, whereas M-FPC estimates the shape of the θ distribution per group freely. Despite of these differences, the three methods, both their single-

group version and multiple-group version, all perform similarly and hence they can be used exchangeably whenever situation allows. Last but not least, in addition to the proposed new item calibration method, the challenge could also be potentially resolved by using a multidimensional IRT (MIRT) calibration. This is because MIRT calibration also takes into account all item responses in the routing block simultaneously. Future studies could compare the MIRT calibration versus the several methods considered herein.

References

Ban, J.-C., Hanson, B. H., Wang, T. Y., Yi, Q., & Harris, D. J. (2001). A comparative study of on-line pretest item-calibration/scaling methods in computerized adaptive testing. *Journal of Educational Measurement, 38*, 191–212.

Beaton, A. E., & Zwick, R. (1992). Overview of the national assessment of educational progress. *Journal of Educational Statistics, 17*, 95-109.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika, 46*, 443-459.

Cai, L. (2008). *A Metropolis-Hastings Robbins-Monro algorithm for maximum likelihood nonlinear latent structure analysis with a comprehensive measurement model*. Unpublished doctoral dissertation, Department of Psychology, University of North Carolina at Chapel Hill.

Cai, L., Roussos, L., & Wang, X. (2018). *Comparison of calibration and drift detection methods under multistage testing*. Paper presented at the NCME annual meeting, New York City, NY.

Cai, L., Yang, J. S., & Hansen, M. (2011). Generalized full-information item bifactor analysis. *Psychological Methods, 16*, 221-248.

Chang, H.-H., & Ying, Z. L. (1996). A global information approach to computerized adaptive testing. *Applied Psychological Measurement, 20*, 213-229.

Chang, H., & Ying, Z. (2008). To weight or not to weight? Balancing influence of initial items in adaptive testing. *Psychometrika*, *73*, 441-450.

Chen, P., & Wang, C. (2016). A new online calibration method for multidimensional computerized adaptive testing. *Psychometrika, 81*, 674–701.

Chen, P., Wang, C., Xin, T., & Chang, H.-H. (2017). Developing new online calibration methods for multidimensional computerized adaptive testing. *British Journal of Mathematical and Statistical Psychology, 70*, 81-117.

Dean, V., Martineau, J. (2012). A state perspective on enhancing assessment and accountability systems through systematic implementation of technology. In Lissitz, R. W., Jiao, H. (Eds.), *Computers and their impact on state assessment: Recent history and predictions for the future* (pp. 55-77). Charlotte, NC: Information Age.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data viathe EM algorithm. *Journal of the Royal Statistical Society, Series B, 39*, 1–38.

Eggen, T. J. H. M., & Verhelst, N. D. (2011). Item calibration in incomplete designs. *Psychológica, 32*, 107-132.

Glas, C. A. W. (2010). Item parameter estimation and item fit analysis. In W. J. van der Linden & C. A. W. Glass (Eds.), *Elements of adaptive testing* (pp. 269–288). NewYork: Springer.

Guo, R., Zheng, Y., & Chang, H.-H. (2015). A stepwise test characteristic curve method to detect item parameter drift. *Journal of Educational Measurement, 52*, 280-300.

Han, K. T., & Guo, F. (2014). Impact of violation of the missing-at-random assumption on full-information maximum likelihood method in multidimensional adaptive testing. *Practical Assessment, Research & Evaluation*, *19*(2). Available online: http://pareonline.net/getvn.asp?v=19&n=2.

Jewsbury, P., & van Rijn, P. (2018). *Random missing in multidimensional multistage testing: the importance of multivariate latent variable models*. Paper presented at the NCME annual meeting, New York City, NY.

Kim, S. (2006). A comparative study of IRT fixed parameter calibration methods. *Journal of Educational Measurement, 43*, 355-381.

Kim, S., & Kolen, M. (2016). Multiple group IRT fixed-parameter estimation for maintaining an established ability scale. Center for Advanced Studies in Measurement and Assessment Report #49, https://education.uiowa.edu/sites/education.uiowa.edu/files/documents/centers/casma/casma-research-report-49.pdf

Lissitz, B., Jiao, H., Li, M., Lee, D., & Kang, Y. (2014). *Software packages for multiple group IRT analysis and accuracy of parameter estimates*. Executive Report for the Maryland State Department of Education.

Little, R.J.A. & Rubin, D.B. (1987). *Statistical Analysis with Missing Data*. New York: John Wiley & Sons.

Liu, O., Wilson, M., & Paek, I. (2008). A multidimensional Rasch analysis of gender differences in PISA Mathematics. *Journal of Applied Measurement, 9*, 18-35.

Lu, R., Jia, Y., & Wu, M. (2018). *Using design information in item parameter estimation with multistage testing*. Paper presented at the NCME annual meeting, New York City, NY.

Lu, R., Jia, Y., & Wu, M. (2017). *Population definition and Identification, priors, and non-random samples*. Paper presented at the NCME annual meeting, San Antonio, TX.

Mislevy, R. J. (1984). Estimating latent distributions. *Psychometrika, 49,* 359–381.

Mislevy, R. J. (1991). Randomization-based inference about latent variables from complex samples. *Psychometrika, 56*, 177-196.

Mislevy, R. J., Beaton, A. E., Kaplan, B., & Sheehan, K. M. (1992). Estimating population characteristics from sparse matrix samples of item responses. *Journal of Educational Measurement, 29*, 133-161.

Mislevy, R.J. & Sheenan, K.M. (1989). The role of collateral information about examinees in item parameter estimation. *Psychometrika*, 54, 661-680.

Mislevy, R.J. & Wu, P-K (1996). Inferring examinee ability when some item responses are missing. *Research Report RR-96-30-ONR*. Princeton: Educational Testing Service.

Rubin, D. B. (1976). Inference and missing data. *Biometrika, 63*, 581-592.

Rubin, D. B. (1991). EM and beyond. *Psychometrika, 56*, 241-254.

Rubin, D. B., & Thayer, D. T. (1982). EM algorithms for ML factor analysis. Psychometrika, 47, 69-76.

Swaminathan, H., & Gifford, J. A. (1986). Bayesian estimation in the three-parameter logistic model. *Psychometrika*, *51*, 589-601.

Thissen, D., & Wainer, H. (1982). Some standard errors in item response theory. *Psychometrika*, *47*, 397-412.

Van Groen, M. (2017). *Multistage testing with multiple subjects*. Invited talk at the International Association of Computerized Adaptive Testing (IACAT), Niigata, Japan.

Wang, C., Zheng, Y., & Chang, H. (2014). Does standard deviation matter? Using "standard deviation" to quantify security of multistage testing. *Psychometrika*, 79, 154-174.

Wainer, H. (1990). *Computerized adaptive testing: A primer*. Hillsdale: Erlbaum.

Weiss, D. J. (1982). Improving measurement quality and efficiency with adaptive testing. *Applied Psychological Measurement, 6*, 473-492.

Woodruff, D. J., & Hanson, B. A. (1996). *Estimation of item response models using the EM*

algorithm for finite mixtures (ACT Research Report 96-6). Iowa City, IA: ACT, Inc.

Wu, M., & Lu, R. (2017). Multi-stage testing simulation studies. Paper presented at the NCME annual meeting, San Antonio, TX.

Wu, M., & Xi, N. (2017). Multi-stage testing in the 2015 NAEP mathematics DBA field trial. Paper presented at the NCME annual meeting, San Antonio, TX.

Yan, D. L., Von Davier, A. A., & Lewis, C. (2014). *Computerized multistage testing: Theory and applications*. NY: CRC press.

Appendix

In this Appendix, we provide derivations showing that when the MAR assumption is satisfied (i.e., routing based on $\hat{\theta}$), the EM algorithm introduced in section 3.2 can also proceed based solely on the observed data.

Specifically, for the E-step, we can write the conditional expectation as follows,

$$E_{(\boldsymbol{y}_{mis},\boldsymbol{\theta})|\boldsymbol{y}_{obs},\boldsymbol{\Delta}^r,\boldsymbol{\pi}^r}\left(\log\left(L(\boldsymbol{\Delta},\boldsymbol{\pi}|\boldsymbol{Y},\boldsymbol{\theta},\boldsymbol{m})\right)\right)$$

$$= E_{(\boldsymbol{y}_{mis},\boldsymbol{\theta})|\boldsymbol{y}_{obs},\boldsymbol{\Delta}^r,\boldsymbol{\pi}^r}\left(\sum_{i=1}^{N}\log\left(L(\boldsymbol{\Delta},\boldsymbol{\pi}|\boldsymbol{y}_{i,obs},\boldsymbol{y}_{i,mis},\theta_i)\times h_{\varphi}\left(\boldsymbol{m}_i\middle|\boldsymbol{y}_{i,obs},\boldsymbol{y}_{i,mis},\theta_i\right)\right)\right)$$

$$= \sum_{i=1}^{N} E_{(\boldsymbol{y}_{i,mis},\theta_i)|\boldsymbol{y}_{i,obs},\boldsymbol{\Delta}^r,\boldsymbol{\pi}^r}\left(\log\left(L\left(\theta_i,\boldsymbol{\Delta}|\boldsymbol{y}_{i,obs}\right)\times L\left(\theta_i,\boldsymbol{\Delta}|\boldsymbol{y}_{i,mis}\right)\times g(\theta_i|\boldsymbol{\pi})\times h\left(\boldsymbol{m}_i\middle|\boldsymbol{y}_{i,obs}^{R}\right)\right)\right)$$

$$= \sum_{i=1}^{N}\sum_{k=1}^{K}\left(\log\left(L\left(\theta_k,\boldsymbol{\Delta}|\boldsymbol{y}_{i,obs}\right)\times\pi_k\times h_{\varphi}\left(\boldsymbol{m}_i\middle|\boldsymbol{y}_{i,obs}^{R}\right)\right)\times p\left(\theta_k\middle|\boldsymbol{y}_{i,obs},\boldsymbol{\Delta}^r,\boldsymbol{\pi}^r\right)\right)$$

$$+\sum_{i=1}^{N}\sum_{k=1}^{K} E_{(\boldsymbol{y}_{i,mis})|\boldsymbol{y}_{i,obs},\boldsymbol{\Delta}^r,\boldsymbol{\pi}^r}\left(\log\left(L\left(\theta_k,\boldsymbol{\Delta}|\boldsymbol{y}_{i,mis}\right)\right)\times p\left(\theta_k\middle|\boldsymbol{y}_{i,obs},\boldsymbol{\Delta}^r,\boldsymbol{\pi}^r\right)\right)$$

$$=\sum_{i=1}^{N}\sum_{k=1}^{K}\left(\log\left(L\left(\theta_k,\boldsymbol{\Delta}|\boldsymbol{y}_{i,obs}\right)\times\pi_k\right)\times p\left(\theta_k|\boldsymbol{y}_{i,obs},\boldsymbol{\Delta}^r,\boldsymbol{\pi}^r\right)\right)+$$

$$\sum_{i=1}^{N}\sum_{k=1}^{K}\left(\log\left(h_\varphi\left(\boldsymbol{m}_i|\boldsymbol{y}_{i,obs}^r\right)\right)\times p\left(\theta_k|\boldsymbol{y}_{i,obs},\boldsymbol{\Delta}^r,\boldsymbol{\pi}^r\right)\right)+$$

$$\sum_{i=1}^{N}\sum_{k=1}^{K}E_{\boldsymbol{y}_{i,mis}|\boldsymbol{y}_{i,obs},\boldsymbol{\Delta}^r,\boldsymbol{\pi}^r}\left(\log\left(L\left(\theta_k,\boldsymbol{\Delta}|\boldsymbol{y}_{i,mis}\right)\right)\times p\left(\theta_k|\boldsymbol{y}_{i,obs},\boldsymbol{\Delta}^r,\boldsymbol{\pi}^r\right)\right) \quad \text{(A1)}$$

The second to the last equality holds because the expectation $E_{(\boldsymbol{y}_{i,mis},\theta_i)|\boldsymbol{y}_{i,obs},\boldsymbol{\Delta}^r,\boldsymbol{\pi}^r}$ is actually a double integral, one with respect to the distribution of $\boldsymbol{y}_{i,mis}$ and the other with respect to the distribution of θ_i. Then, the first term in this equality is irrelevant to $\boldsymbol{y}_{i,mis}$ hence it can be taken outside the expectation with respect to $\boldsymbol{y}_{i,mis}$, resulting in only one integral that is written as a numeric sum over *K*.

In the last equality in Eq. (A1), the first term is simply the conditional expectation of the log-likelihood based on *observed* data (i.e., the same as Eq. 5), the second term is irrelevant to the target parameters, whereas the third term actually vanishes in the M-step. The explanation is as follows. Without loss of generality, take item *j* as an example. Take a first-order derivative with respect to Δ_j, we have

$$E_{\boldsymbol{y}_{i,mis}|\boldsymbol{y}_{i,obs},\boldsymbol{\Delta}^r,\boldsymbol{\pi}^r}\left(\frac{\partial\log[L(\theta_k,\boldsymbol{\Delta}|\boldsymbol{y}_{i,mis})]}{\partial\Delta_j}p\left(\theta_k|\boldsymbol{y}_{i,obs},\boldsymbol{\Delta}^r,\boldsymbol{\pi}^r\right)\right)=p\left(\theta_k|\boldsymbol{y}_{i,obs},\boldsymbol{\Delta}^r,\boldsymbol{\pi}^r\right)\times$$

$$E_{\boldsymbol{y}_{i-j,mis}}\left[E_{y_{ij,mis}}\left(\frac{\partial\log[L(\theta_k,\boldsymbol{\Delta}|y_{ij,mis})]}{\partial\Delta_j}\right)\right] \quad \text{(A2)}$$

where $y_{ij,mis}$ denotes the missing responses of person *i* on item *j*, and $\boldsymbol{y}_{i-j,mis}$ denotes the missing responses of person *i* on the remaining items except item *j*. Equation (A2) holds because (1) $p\left(\theta_k|\boldsymbol{y}_{i,obs},\boldsymbol{\Delta}^r,\boldsymbol{\pi}^r\right)$ is irrelevant to the distribution of $\boldsymbol{y}_{i,mis}$ and hence it can be taken outside the expectation; and (2) due to the discreetness of $\boldsymbol{y}_{i,mis}$, the expectation with respect to the posterior distribution of $\boldsymbol{y}_{i,mis}$ can be expanded as a series of expectations.

Because $L\left(\theta_k, \boldsymbol{\Delta}\middle|y_{ij,mis}\right) = P\left(y_{ij,mis} = 1\middle|\theta_k, \boldsymbol{\Delta}\right)^{y_{ij,mis}}(1 - P\left(y_{ij,mis} = 1\middle|\theta_k, \boldsymbol{\Delta}\right))^{1-y_{ij,mis}}$

Consider item parameter a_j as an example, then

$$\frac{\partial \log[L(\theta_k,\boldsymbol{\Delta}|y_{ij,mis})]}{\partial \mathrm{a}_j} = 1.7\left(y_{ij,mis} - P\left(y_{ij,mis} = 1\middle|\theta_k, \boldsymbol{\Delta}\right)\right). \quad \text{(A3)}$$

And because $y_{ij,mis}$ follows a Bernoulli distribution, it is easily shown that the expectation of (A3) with respect to the distribution of $y_{ij,mis}$, i.e., $E_{y_{ij,mis}}\left(\frac{\partial \log[L(\theta_k,\boldsymbol{\Delta}|y_{ij,mis})]}{\partial \Delta_j}\right)$=0. As a result, Eq. (A2) also becomes 0 and hence it vanishes in the M-step. Therefore, with missing data satisfying MAR, the EM algorithm can proceed in the same fashion as in section 3.2 using the observed data.