Research Report

Practical Considerations in Item Calibration With Small Samples Under Multistage Test Design: A Case Study

ETS RR–24-03

Hongwen Guo
Matthew S. Johnson
Daniel F. McCaffrey
Lixong Gu

December 2024

ETS Research Report Series

Since its 1947 founding, ETS has conducted and disseminated scientific research to support its products and services, and to advance the measurement and education fields. In keeping with these goals, ETS is committed to making its research freely available to the professional community and to the general public. Published accounts of ETS research, including papers in the ETS Research Report series, undergo a formal peer-review process by ETS staff to ensure that they meet established scientific and professional standards. All such ETS-conducted peer reviews are in addition to any reviews that outside organizations may provide as part of their own publication processes. Peer review notwithstanding, the positions expressed in the ETS Research Report series and other published accounts of ETS research are those of the authors and not necessarily those of the Officers and Trustees of Educational Testing Service.

The Daniel Eignor Editorship is named in honor of Dr. Daniel R. Eignor, who from 2001 until 2011 served the Research and Development division as Editor for the ETS Research Report series. The Eignor Editorship has been created to recognize the pivotal leadership role that Dr. Eignor played in the research publication process at ETS.

RESEARCH REPORT

Practical Considerations in Item Calibration With Small Samples Under Multistage Test Design: A Case Study

Hongwen Guo, Matthew S. Johnson, Daniel F. McCaffrey, & Lixong Gu

ETS, Princeton, New Jersey, United States

The multistage testing (MST) design has been gaining attention and popularity in educational assessments. For testing programs that have small test-taker samples, it is challenging to calibrate new items to replenish the item pool. In the current research, we used the item pools from an operational MST program to illustrate how research studies can be built upon literature and program-specific data to help to fill the gaps between research and practice and to make sound psychometric decisions to address the small-sample issues. The studies included choice of item calibration methods, data collection designs to increase sample sizes, and item response theory models in producing the score conversion tables. Our results showed that, with small samples, the fixed parameter calibration (FIPC) method performed consistently the best for calibrating new items, compared to the traditional separate-calibration with scaling method and a new approach of a calibration method based on the minimum discriminant information adjustment. In addition, the concurrent FIPC calibration with data from multiple administrations also improved parameter estimation of new items. However, because of the program-specific settings, a simpler model may not improve current practice when the sample size was small and when the initial item pools were well-calibrated using a two-parameter logistic model with a large field trial data.

Keywords MST; FIPC; item scaling; data matching

doi:10.1002/ets2.12376

The multistage testing (MST) design has been gaining attention and popularity in national and international educational assessments such as the National Assessment of Educational Progress (NAEP), Programme for International Student Assessment (PISA), Trends in International Mathematics and Science Study (TIMSS), and Progress in International Reading Literacy Study (PIRLS) (see Jewsbury & van Rijn, 2020; Mead, 2006; Mullis & Martin, 2019; von Davier et al., 2006; Yamamoto et al., 2018). The MST design has several advantages: An MST test can provide accurate ability estimates similar to computer adaptive testing (CAT); that is, it can provide more accurate ability estimation with fewer items than the number of items required by a linear test because the MST test adapts to a test taker's ability. An MST test enables test developers control of each test form in content and structure, and it is friendly to test takers because they can revisit the items within each item block (Hambleton & Xing, 2006; Wainer & Mislevy, 1990; Wang et al., 2020). More importantly, through matching overall item difficulties to a target population and shortened tests, test takers might be more engaged with an MST test, better enabling the collection of valid and effortful responses (Ercikan et al., 2020; Guo & Ercikan, 2021; Wise, 2021).

In a typical MST design, the routing block in the first stage contains items with a wide spread of item difficulty, acting as a linear test for all test takers. In the next stage of MST, item blocks (also called modules or target blocks) consisting of easy, medium, and hard items in difficulty are administered. Based on their performance on the routing block, test takers are routed to a target block that is most appropriate for their abilities. For subsequent stages, test takers are routed to a suitable target block in the next stage based on their performance on previous stages, and the commonly used MST designs in large-scale assessments have two or three stages. As in any CAT and MST tests, new items have to be developed and calibrated to replenish the item pools to assemble new test forms. In practice, the new items are often assembled in one item block (called the pretest item block), embedded in the operational forms, and administrated to test takers. The pretest item block is assembled in a way similar to an operational item block (i.e., routing block) in terms of the number of items and content but with unknown item parameters; this design helps to disguise the pretest items and avoid test takers skipping the pretest blocks so that new items can be appropriately calibrated.

Corresponding author: H. Guo, E-mail: hguo@ets.org

When the sample sizes of test takers are relatively large, pretest items can be calibrated, separately or concurrently, with operational items using item response theory (IRT) models (Kolen & Brennan, 2004). However, it is challenging to calibrate new items accurately with small samples of test takers in a linear test form (Drasgow, 1989; Seong, 1990; Stone, 1992), let alone items that are administered under the MST design because of missing data by design. Small samples of test takers may occur for many reasons. In highly specialized occupations, the number of people working in a field may be limited, thus the volume of test takers for any given test administration may be small. In addition, tests are often administered multiple times per year, which further reduces the number of test takers in any given administration (Peabody, 2020). Various practical challenges may also lead to smaller samples for existing testing programs (Guo, 2022; Jiao & Lissitz, 2020), such as when the frequencies of administrations have to be increased to meet test takers' needs (convenience and flexibility of testing dates, for example) or to control item exposure (because of remote testing, for example), or when the test volumes drop because of unexpected factors (the test optional policy, for example).

In the current research, we used the item pools from an operational MST program to illustrate how research studies can be built upon literature and program-specific data to help fill the gaps between research and practice and to make sound psychometric decisions to address the small-sample issues. The sample size in our research was the total number of test takers on a specific test form. The three studies in our research included choice of item calibration methods, choice of data collection designs to increase sample sizes, and choice of IRT models in producing the score conversion tables. Our studies were conducted to answer the following three research questions and fill the gaps between program practices and research findings with easy operational implementation and sound psychometric properties as priorities. Under the MST design, the research questions we explored were:

Q1: Which calibration method should be recommended to calibrate new items in the small-sample situation? How small can the samples be?
Q2: What are alternative ways in data collection design to increase the sample sizes?
Q3: Will a simpler IRT model be justified in the small-sample situation for the studied testing program in terms of score reporting?

In the following section, we review relevant literature on IRT calibrations on a linear form or under an MST design. To illustrate various practical considerations in psychometric decisions, we used a newly established testing program as an example in the current research. In the Study Design section, we introduce the MST design of the studied program and its routing decision. We also specify the factors we manipulated in simulations when some model assumptions may be violated.

In the Method section, we describe three calibration methods in detail: One is commonly used in testing program practices, one is commonly used in research studies, and one is proposed in the current research to address the mismatch between the assumed ability distribution and empirical data. In the Results section, using simulated data, we first present estimation errors in item parameters that resulted in the first two studies using the two-parameter logistic (2PL) IRT calibration methods. We then report conversion table differences if the simpler one-parameter logistic (1PL) model was used in item calibration in the third study. In the last section, we discuss the findings, limitations, and recommendations. Even though the current research used specific program data, the procedures, methods, and considerations for making program decisions are generalizable to other testing programs to solve program-specific challenges.

Literature Review

In this section, we review IRT item calibration studies from different but interconnected perspectives: calibration with small samples on a linear form, item scaling and fixed parameter calibration on a linear form, item calibration consideration under the MST design, and the 2PL versus 1PL model choice. At the end of each subsection, we highlight the issues that need further investigations for operational MST testing programs with specific constraints and the added values the current research may contribute to literature and program practice.

IRT Calibration With Small Samples

In a typical IRT item calibration, it is generally assumed that the latent ability/trait (denoted by θ) is normally distributed in the population when estimating the logistic IRT model parameters (de Ayala, 2009; Reise et al., 2018). When this normality assumption is violated, models are misspecified, and item and person parameter estimates are inaccurate, theoretically. If either the logistic response model or the assumed ability distribution is incorrect, the statistical properties of marginal maximum likelihood estimates (MMLEs) — the most commonly used estimation method — may not hold (Mislevy & Sheehan, 1989).

Using the BILOG software package, Seong (1990) showed that, when the underlying θ and prior θ distributions were matched, the 2PL IRT item discrimination and difficulty parameters were estimated more accurately as the sample size increased for a test of 45 items. On the other hand, when the underlying θ and prior θ distributions were mismatched, appropriate specification of the prior distribution in the calibration algorithm increased the accuracy of theta and item parameter estimation for large sample sizes ($N = 1{,}000$), but appropriate specification did not help much with item parameter estimation for the small sample size ($N = 100$). Using the BILOG and MULTILOG software packages, Stone (1992) found in simulations that, under varying θ distributional assumptions in the 2PL IRT models, MMLEs of item difficulty parameters (the b parameters) were generally precise and stable in small samples. When the true θ distribution was normal, MMLEs of item discrimination parameters (the a parameters) were also generally precise and stable.

In his book, de Ayala (2009, p. 105) reviewed many previous studies on the accuracy of the 2PL IRT parameter estimation and pointed out that, "assuming MMLE, the use of a prior distribution for a, and favorable conditions (e.g., θ/prior distribution match, etc.), it appears that a calibration sample size of at least 500 persons and instruments of 20 or more items tend to produce reasonably accurate item parameter estimates."

However, in practice, the underlying θ distribution is unknown, and the "favorable conditions" are hard to verify. In addition, the above studies used linear test forms. Thus, in the current research, we simulated MST data with different symmetric and asymmetric θ distributions in Study 1, compared to different calibration approaches and their robustness for different sample sizes.

Item Scaling and Fixed Item Parameter Calibration

When calibrating new items with different test-taker samples, the item parameter estimates from separate calibrations are on different scales, and thus the estimates need to be transformed to the item pool scale through common/anchor items. Because of the invariance of the 2PL IRT model under a linear transformation of θ, various approaches have been proposed to find the coefficients in the linear transformation, among which the Stocking and Lord (SL) approach (Stocking & Lord, 1983) is the most commonly used in research and practice (Kim & Kolen, 2019; Kolen & Brennan, 2004). The SL approach estimates the linear coefficients by minimizing the squared difference between the test characteristic curves (TCCs) over the anchor items.

Alternatively, the fixed item parameter calibration (FIPC) approaches were proposed to calibrate new/pretest items, which do not require the item scaling step (Ban et al., 2006; Chen et al., 2017; Wainer & Mislevy, 1990). In FIPC, one fixes the item parameters from the pool and calibrates only the new items. Kim (2006) compared five FIPC methods on simulated linear test forms under the unidimensional IRT models and using different software packages (BILOG, ICL, and PARSCALE). Simulation results showed that new item parameter estimates were increasingly accurate with increased sample sizes (from $N = 300$ to $N = 3{,}000$), and the accuracy was slightly improved with more fixed items (from 10 items to 40 items). Among the five FIPC methods, the multiple weights updating and multiple expectation – maximization (EM) cycles (MWU-MEM) method appeared to perform properly and robustly under different θ distributions. Ban et al. (2006) compared several fixed ability estimate methods and FIPCs under the unidimensional IRT models with different sample sizes (300, 1,000, and 3,000), and they found that the MWU-MEM methods appeared to be the best choice. For multidimensional IRT models, there were mixed findings between fixing ability estimates and fixing item parameter approaches (Chen et al., 2017).

The feasibility of the MWU-MEM method has been supported in several subsequent studies (e.g., DeMars & Jurich, 2012; Kim & Kolen, 2019; König et al., 2021). For example, Kim and Kolen (2019) applied the MWU-MEM FIPC approach to multiple-group test data on linear test forms. Using the ICL software package, they investigated multigroup item calibration under three different linking design for linear test forms of 40 items with 10 anchor items. Comparison between FIPC and traditional separate calibration with scaling showed that, with sample sizes of 500 and 2,000, respectively, the multigroup FIPC method performed nearly equally to or better than the traditional approach in recovering the underlying ability distributions and the new item parameters.

However, most of the mentioned studies assumed symmetric (i.e., normal) ability distributions and item difficulties. In the current research, we tried to address to what extent one method performed better than the others for skewed ability distributions with the program-specific item pool and how small a sample size the program could consider to maintain the item pool quality. In addition, we introduced a new approach, the MDIA-based approach, to address the mismatch between the model assumption and empirical data on the ability distributions (refer to the Method section for details).

MST Item Calibrations

Under an MST design, after the first stage (the routing block/module), test takers are split into separate subgroups by ability (most commonly three subgroups) and routed to different target blocks/modules. Hence, there are missing data by design. Because whether a response is missing or not depends only on the observed responses under an MST design, Mislevy and Wu (1988, 1996) argued that the missing at random (MAR) assumption held and missing data could be ignored when making inference about θ. Eggen and Verhelst (2011) provided justification for using MMLE in the MST item calibration. Wang et al. (2020) further showed that, in practice when routing is based on the estimated θ, the MAR held, and the single group calibration (i.e., calibrate all items together in a single group) produced unbiased item parameter estimates for unidimensional IRT models. On the other hand, concurrent multigroup calibration can be used only when the true θ is known. Wang et al. discussed the three item calibration methods (MMLE, EM, FIPC) in the context of MST in their study, and the single group FIPC (calibrating routing block items first and then fixing them to calibrate target block items) performed the best in the large sample ($N = 3{,}000$) simulations. Wang et al. also suggested, when there were subscales, item calibration for each subscale should include items used for routing decision. Jewsbury and van Rijn (2020) further considered MST item calibrations for multidimensional IRT models, where the routing decision was based on performance on items from multiple subscales. Their findings suggested that multidimensional IRT should be used to estimate the item parameters for each scale simultaneously with all items used for routing decision regardless of subscales.

Given that none of the above studies focused on new item calibrations with small samples under the MST design, the current research may constitute a contribution to the MST literature and practice at large.

IRT Model

As shown in aforementioned studies, it is challenging to estimate the item discrimination parameter a accurately for a small sample of test takers.

O'Neill et al. (2020) evaluated how sample sizes affected the stability of item calibrations and person ability estimates in a Rasch model. They used a resampling design to create varying sample size conditions on a long linear test (with $J = 240$ items), and their results empirically demonstrated that even imprecise calibrations could occasionally be used for the purposes of equating without much damage to the person ability estimate when there were a reasonable number of anchor items with precise and stable calibrations.

Using numerical approximations of estimation error, Lord (1983) showed that, for small samples under certain situations, it was better to use the Rasch model even though the Rasch model was incorrect. Under his studied test form conditions (5-, 10-, or 15-item forms constructed from a 50-item vocabulary test), the observed raw sum score x under the 1PL model may be slightly superior to the weighed sum score $\sum \hat{a}_i x_i$ under the 2PL model estimator as an approximation of the true scores on f ive θ points $(-2, -1, 0, 1, 2)$ in terms of estimation errors when the sample size was less than 100 or 200. de Ayala (2009, p. 152) summarized model selection studies, which indicated that the Rasch model yielded rather good estimates of ability when the discrimination parameters could be assumed to be roughly equal or when the number of items was very large.

However, for an MST test, the test length is usually much shorter, and the discrimination parameters may not be similar. It is unclear whether the findings could be generalized for a specific testing program. In addition, for testing programs that use the IRT true score equating under the MST design, estimated abilities and their transformations are not used in score reporting. Instead, the observed score, which is linked to a reported score through a conversion table, is used. The conversion table maps a observed score to the reported score scale through the TCC (Kolen & Brennan, 2004). The estimation error of item parameters (and ability) is collectively and indirectly reflected in the test form's conversion table. Therefore, in the current research, instead of evaluating the estimation error of ability or true scores, we focused on changes in conversion tables when different IRT models were used to produce TCCs with different sample sizes. We evaluated whether there was enough evidence in the variations of the conversion tables produced by true score equating with 1PL and 2PL models, so that a recommendation could be suggested to the testing program on whether to switch to the 1PL model, instead of the current practice of using the 2PL model.

Study Design

Because the goal of the current research is to provide guidance on operational practice, easy implementation of the recommended procedures is a priority.

The current research contains three simulation studies. Study 1 attempts to address general issues in calibration methods under small samples. We compared three 2PL calibration methods for the new items in the pretest item block. Study 1 focused on the magnitude of accuracy improvement under different conditions by using FIPC methods and comparing them to the traditional method (the separate calibration with scaling method; the default calibration method for the studied program). In addition, we proposed and experimented with the MDIA-based approach for item calibrations (refer to the Method section for more details). Study 2 was more program-specific and focused on data collection schemes to increase sample sizes for new item calibration. We compared item estimation errors in the one-step 2PL FIPC and the two-step 2PL FIPC with additional data, in case the sample sizes were small in one test administration. Study 3 was also program-specific. Using the FIPC calibration method, we compared the conversion tables produced by both a 1PL model and a 2PL model to evaluate whether the testing program could consider the simpler 1PL model in the small-sample situation.

In the three studies, we assumed that the items in the initial pool were accurately estimated in a 2PL IRT calibration because of the large sample size of test takers in a field test.

MST Design

The studied MST design had two stages, as shown in Figure 1. The first stage consisted of a routing block. Based on their performance on the first stage, a test taker was routed to one of three target blocks in the second stage: an easy, medium difficult, or difficult item block.

In addition, all test takers took a pretest item block similar to the routing block in terms of the number of items and the content coverage.

Underlying Ability Distribution

Given the assumption of the θ distribution being $N(0, 1)$ in most IRT software when using MMLE algorithms and the importance of the prior ability distribution in calibration, we used the five underlying θ distributions in simulation to investigate item parameter estimation accuracy.

The first two were skewed distributions, which might occur when the test was too easy or too hard for the test-taker population (with the ceiling or flooring effect). As in Seong (1990), we used a chi-square distribution having eight degrees of freedom (df, the skewness was 1) as the positively skewed (PS) distribution. The negatively skewed (NS) distribution was the mirror image of the previous one (refer to Figure 2). We also considered three symmetric/normal distributions, $N(0, 1)$, $N\left(0.5{,}1.2^2\right)$, and $N\left(1{,}1.4^2\right)$, as was used in Kim (2006). The later two normal distributions were to mimic test population shift from a standard normal distribution.

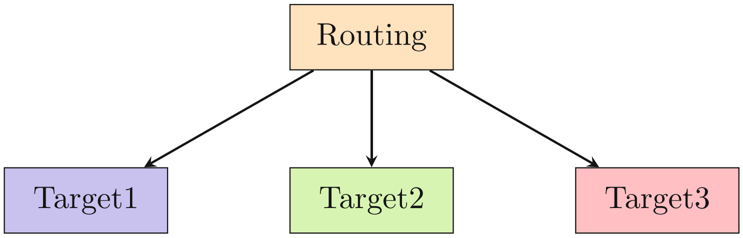

Figure 1 The MST design

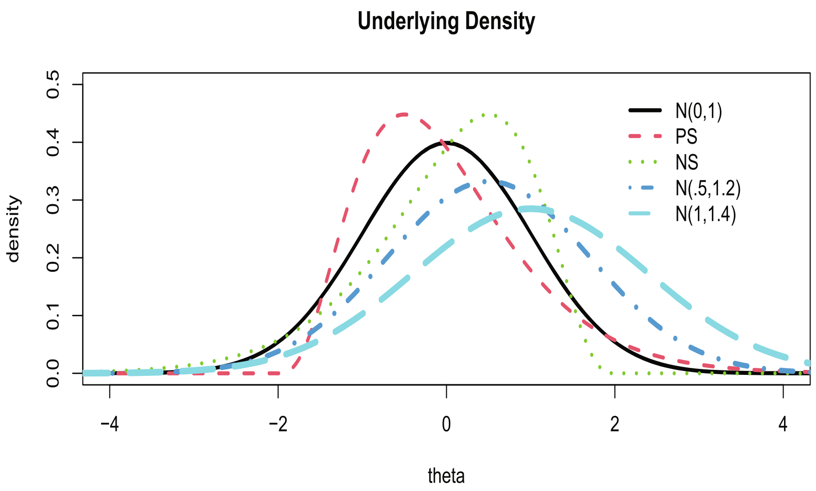

Figure 2 The θ distributions. PS = positively skewed; NS = negatively skewed

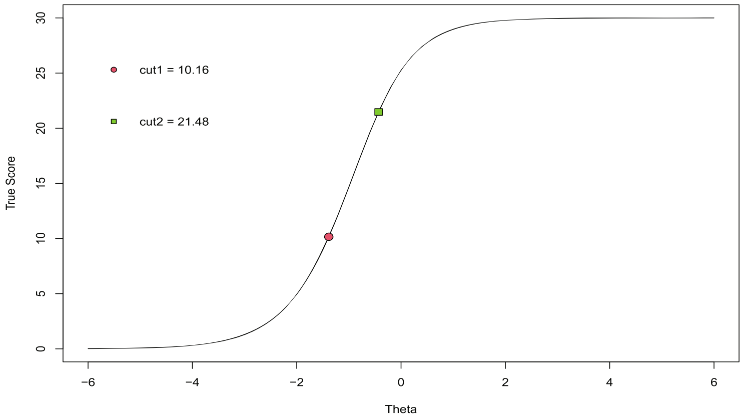

Figure 3 Test characteristic curve and routing score cuts on the original base form of 30 items

Routing Decision

The routing decision uses the defined population intervals method (Luecht et al., 2006), which specified the relative proportions of test takers in the population expected to be routed to each of the three target blocks. More specifically, in the studies, the routing decision was based on two cut points on the base form, so that about one third of the test takers would be routed to the easy, medium difficult, and difficult target blocks, respectively.

We used the two scale score points (cut1 = 10.16 and cut2 = 21.48) produced by the original base form of 30 items (refer to Figure 3). Using the Newton-Raphson procedure, we identified the corresponding θ_1 and θ_2 based on the TCC from a 2PL model. In the following simulations, test takers' preliminary abilities were estimated on the routing block, and those with estimated ability $\widehat{\theta}$ larger (smaller) than θ_2 (θ_1) would be routed to the difficult (easy) target block. Those with $\theta_1 < \widehat{\theta} < \theta_2$ would be routed to the medium difficult target block.

Three Studies

Study 1

Based on the literature review, the calibration methods in Study 1 were chosen to be the traditional separate calibrations with SL scaling commonly used in practice (Kolen & Brennan, 2004) and FIPC by fixing the operational item parameters, commonly used in research (Kim, 2006). We also introduced a new calibration approach with the MDIA-matching method (Haberman, 2009). Please refer to the Method section for details.

In the simulations, a few factors were manipulated: the underlying θ distribution (refer to Figure 2), the number of items J in a block/module, and the sample size N. The number of items J in a block/module was chosen to be $J = 15$ for a short test of 30 items or $J = 30$ for a relative long test of 60 items, both of which were seen in practice. The sample size N was chosen to be around 500 for a 2PL IRT calibration (de Ayala, 2009); that is, $N = 250$, 500, or 1,000. In the simulations, data were generated from the 2PL IRT models. Items were drawn from a pool having mean (standard deviation) of 1.00 (0.50) for the a parameter and -0.70 (0.70) for the b parameter to mimic the studied program and to draw relatively more general insights.

Study 2

In Study 2, we focused only on FIPCs based on findings from Study 1 to address program-specific issues. We compared a one-administration FIPC with initial data alone to a two-administration FIPC that combined data collected from two test administrations containing the same new items in order to evaluate the magnitude of improvement in parameter estimation for new items with different data collection schemes.

In the simulations, items were drawn from the initial item pool of the operational testing program, with mean (standard deviation) of 0.92(0.43) for the a parameter and $-1.18(0.84)$ for the b parameter. The factors that were manipulated in simulations were the θ distribution and sample size N, while the block length J was fixed as 15 as in the testing program (i.e., the whole test length was 30). Because the testing program found that test takers were highly competent, the positively skewed θ distribution was not considered.

One-administration FIPC was conducted as in Study 1. For two-administration FIPC, we assumed that the new items were initially assembled in the pretest block on Test Form 1, which was first administrated (Admin 1) to a sample of test takers (size $= N_1$). In the second administration, Test Form 2 was administrated with the same new items to a different sample of test takers (size $= N_2$). With data from the two administrations, a concurrent FIPC calibration (size $= N_1 + N_2$) was conducted for these new items, where operational item parameters were fixed.

There were several ways to embed these new items in Form 2, as shown in Figure 4. In Case A, the intact pretest block of the new items on Form 1 (shaded gray) were administrated again as a pretest block on Form 2. Items in the routing block (shaded yellow) on Form 2 were assumed to be from the item pool, so that routing decision was assumed to be accurate. This design was to maximize the sample size for the concurrent calibration of the new items.

In Case B, the newly calibrated items (still shaded gray) were assembled into the target blocks on Test Form 2 according to their initial FIPC estimates on Form 1. The remaining items in the target blocks (shaded blue) and the routing block on Form 2 (shaded yellow) were from the item pool. This way, the pretest block on Form 2 (shaded green) could be used to test new items. This design was to maximize the opportunity of pretesting more new items on Form 2 and ensure an accurate routing decision at the same time. In Case C, the pretested new items on Form 1 were used in the routing block on Form 2 to maximize the sample size for concurrent calibration of these new items and to maximize the opportunity of pretesting more new items on Form 2. Case C involved how inaccurate item parameters might compromise the routing decision, and it warrants further studies. So we focused on the first two cases in Study 2. However, results from Case A would always be better than Case B because of the larger sample size, and subsequently, simulations were conducted only under Case B.

Study 3

When sample sizes became small, Study 3 evaluated whether there was evidence to recommend that the testing program transition to a 1PL model, instead of the current operational procedure of using a 2PL model, while maintaining the stability of reported scores. The program used the 2PL IRT true score equating method to produce a conversion table that maps a true score to a reported score (refer to Figure 5).

Using the Newton Raphson method and the new form TCC, we can find the corresponding θ for a given true score T, and then using this θ and the base/reference form TCC, we found the reported score on the reference form (Guo & Dorans, 2019, 2020; Kolen & Brennan, 2004). For simplicity, the base form raw score scale was used for reporting without further scaling.

In Study 3, we made use of Case C in Figure 4 to evaluate the impact of using pretested items in the routing block on the conversion tables when the item parameters were calibrated with different IRT models and sample sizes. The pretested item parameters were calibrated using the FIPC method as in Study 2 (the one-administration case), using either 1PL or 2PL models.

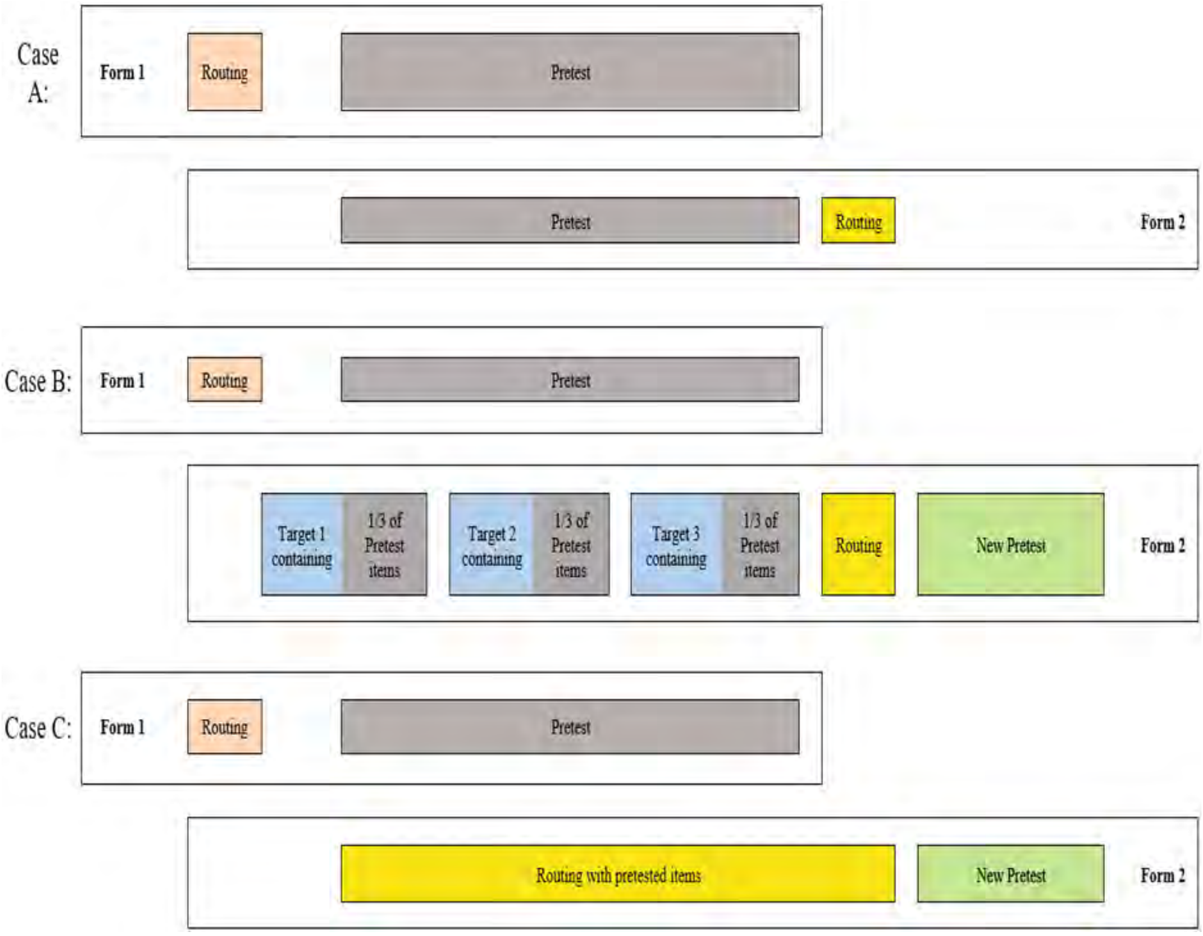

Figure 4 Concurrent calibration designs

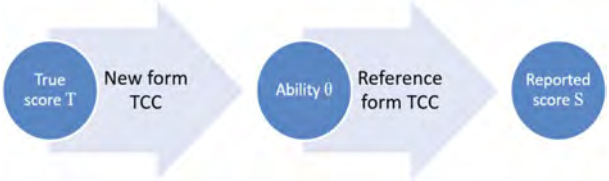

Figure 5 IRT true score conversion table

Evaluation Criterion

Simulations under each condition in the studies were replicated R times. In the first two studies, the mean absolute error (MAE) between the estimated parameter and the true parameter used in the simulation was calculated for each item, and it was then averaged across J items in the item block and across the R replications. That is, let ξ_j be the true item parameter for item j and $\widehat{\xi}_{jr}$ be the estimate in the rth simulation. T he average MAE on the test form of J items was

$$MAE = \frac{1}{R}\sum_{r=1}^{R}\frac{\sum_{j=1}^{J} \mid \widehat{\xi}_{jr} - \xi_j \mid}{J}.$$

Similarly, in Study 3, the conversion table produced by the true item parameters was used as the criterion. The MAE between the criterion conversion table and those using estimated parameters was calculated and then averaged across the R replications for each integer in the raw score range. (MAE is a more robust measure than root mean square error.)

Method

In this section, we briefly discuss calibration methods, item scaling methods, and the MDIA-based sample matching method. More details can be found in the cited references.

Calibration Methods

The 2PL IRT model was used throughout the simulations. In the 2PL model, the item response function for item j is defined as

$$P_j(\theta) = \frac{\exp\left(Da_j\left(\theta - b_j\right)\right)}{1 + \exp\left(Da_j\left(\theta - b_j\right)\right)},$$

where a_j and b_j are the item discrimination and difficulty parameters, $D = 1.702$, and θ is the latent ability variable (Lord, 1980).

Let $\xi = (a, b)$, y_{ij} be the item response to item j from test taker i, and $Y_i = \left(y_{ij}\right)$ the item response vector for test taker i on a test of J items for $i = \{1, 2, \cdots, N\}$. The most commonly used calibrations start from the probability of Y_i for test taker of ability θ_i:

$$P\left(Y_i|\theta_i, \xi\right) = \Pi_j^J \left[P_j\left(\theta_i\right)^{y_{ij}}\left(1 - P_j\left(\theta_i\right)\right)^{1-y_{ij}}\right]. \tag{1}$$

Let $P\left(Y_i\right) = \int P\left(Y_i|\theta_i, \xi\right) g\left(\theta_i|\tau\right) d\theta_i$, where τ is the hyperparameter of the prior ability distribution $g(\cdot)$ (Baker & Kim, 2004; de Ayala, 2009). For the entire data matrix Y, the marginal likelihood function is

$$L(Y) = \Pi_i^N P\left(Y_i\right), \tag{2}$$

and the logarithm of L is

$$\log L = \sum_i^N \log P\left(Y_i\right). \tag{3}$$

In the following subsections, we focus on the separate calibration method and the FIPC method, and then we introduce a calibration method with MDIA matching. Note that the choice of FIPC is based on findings from literature (discussed in Ban et al., 2006, and Chen et al., 2017). In addition, when the item pool is accurately calibrated, FIPC may be conceptually advantageous over the fixed-ability estimate methods because of avoidance of measurement error, particularly for short tests and relatively easy implementation.

Separate Calibrations

In a separate item calibration, the MMLE/expectation-maximization (EM) approach by Bock and Aitkin (1981) is often used, which yields a solution for item parameter estimation that is computationally feasible and consistent under the assumption that the population distribution is known or is concurrently estimated with the correct specification (Baker & Kim, 2004; Dempster et al., 1977; Mislevy & Bock, 1986). The EM algorithm uses an iterative procedure for finding the maximum likelihood estimates of parameters in the presence of the unobserved ξ variables. In the tth step, the EM algorithm alternates between

E-step: Compute $E\left[\log L(Y, \theta|\xi)|Y, \xi_t\right]$ with respect to θ.
M-step: Choose ξ_{t+1} such that the posterior expectation is maximized with respect to ξ.

The process is repeated until a convergence criterion is met. It is usually assumed that there is one population from which the sample test takers are drawn. However, in the EM algorithm, subjects are randomly sampled from $g(\theta|\tau)$, which is assumed to be $N(0, 1)$ in most commonly used IRT packages, even though the EM algorithm also allows an arbitrary distribution of θ in the population sampled (Baker & Kim, 2004; Mislevy & Bock, 1986).

FIPC

FIPC can be viewed as a version of the MMLE/EM method, adapted by fixing a subset of item parameters from the item pool; this way, when new items are calibrated with operational items from the item pool, the new item parameters are automatically put on the fixed scale (Kim, 2006; Wang et al., 2020).

In FIPC, the latent θ distribution is represented by a K-tuple vector $\pi = \left(\pi_1, \cdots, \pi_k\right)$ at K points $\left(q_1, \cdots, q_K\right)$ instead of a continuous distribution as described for the MMLE/EM method. Because of the indeterminacy of the IRT ability scale, there are different ways to update and scale provisional estimates of the underlying ability distribution and item parameters during the EM cycles in FIPC (Kim, 2006; Woodruff & Hanson, 1996).

In the recommended MWU-MEM method (Ban et al., 2006; Kim, 2006), at the tth-step of the iterative EM procedure, the E-step is expressed as

$$\sum_{k=1}^{K} \left[\log L\left(Y_{New} | q_k, \xi_{New}\right) p\left(q_k | Y_{Old}, \xi_{Old}, \xi_{New}^{t}, \pi^{t}\right)\right], \tag{4}$$

where ξ_{Old} are the fixed operational item parameters, and Y_{Old} are the item responses to the operational items. The parameter ξ_{New} is the one that maximizes Equation 4.

SL Item Scaling

As already mentioned, when items are calibrated freely with samples S and T, item parameters need to be put on the same scale through a linear transformation. The SL method (Kolen & Brennan, 2004; Stocking & Lord, 1983) finds a linear transformation (with coefficients A and B) by minimizing the sum of squared differences between two TCCs over the common item set V. That is, A and B are obtained by minimizing

$$\text{SLdiff} = \sum_{i} \text{SLdiff}\left(\theta_i\right),$$

where

$$\text{SLdiff}\left(\theta_i\right) = \left\{\sum_{j\in V} p_{ij}\left(\theta_{S_i} : a_{S_j} b_{S_j}\right) - \sum_{j\in V} p_{ij}\left(\theta_{S_i} : \frac{a_{T_j}}{A}, Ab_{T_j} + B\right)\right\}^2,$$

and $\left(a_{S_j}, b_{S_j}\right)$ and $\left(a_{T_j}, b_{T_j}\right)$ are item parameters estimated separately from the two samples S and T. An iterative approach is used to obtain A and B.

MDIA Matching

Besides these calibration approaches, we also experimented with MDIA (Haberman, 2015), to create a "favorable condition" (de Ayala, 2009), such that the adjusted samples may have a underlying θ distribution matching the prior distribution in the MMLE/EM IRT calibration. The MDIA approach assigns individual weights to individual test takers in the smaller and newer sample, so that the weighed sample is pseudo-equivalent to a larger reference sample in terms of item responses. As in Haberman (2015), let z_i be the considered matching variable (in our case, it is related to the item response vector) for each test taker. The target mean $\overline{Z}$ from the reference sample is

$$\overline{Z} = \sum_{i=1}^{N_R} \frac{z_{iR}}{N_R}, \tag{5}$$

where z_{iR} and N_R are the matching variable vector and the number of test takers in the reference sample, respectively. Using the MDIA approach, the weight w_i is obtained and assigned to test taker i in the new sample, so that

$$\sum_{i=1}^{N} w_i z_i / N = \overline{Z}$$

holds, where $w_i > 0$ and $\sum_i w_i = 1$. The MDIA approach uses the Newton-Raphson method to obtain the individual weight (Haberman, 1984).

Results

All analyses and calibrations were conducted using the R programming language and the associated R packages. The R codes for the key steps of implementation can be found in Appendix A. For each studied condition, the simulation was replicated 100 times.

Note that the routing block should be included in item calibration, following the findings from literature. Based on our exploration, we found that including target blocks produced large errors because of smaller samples in the second stage. Hence, we used the routing block only (labeled as “operational items" in the subsequent tables) in all the three calibration methods. That is, in separate calibrations with SL scaling, the routing block was used as the common item; in calibrations with MDIA, item responses on the routing block were used for matching; and in FIPCs, the routing item parameters were fixed.

Study 1: Three Calibration Methods for New Items

Study 1 compared three calibration methods for the new items: SL scaling, MDIA, and FIPC. In all the calibrations, the reference samples were simulated from a $N(0, 1)$ distribution with a sample size of $N_0 = 5{,}000$. In calibrations with MDIA, we used item responses on the routing block and the first 30 PCA factors from item pairs for matching. (PCA is used to avoid singularity when raw item pairs are employed. Generally, the first 20 to 30 PCA factors explain more than 90% variation in the item pairs.)

The average MAE of item parameter estimates are reported in Tables 1 to 3 for different sample sizes, 250, 500, and 1,000, for the new item calibration.

When the sample size was 250, Table 1 shows that, compared to calibration with SL scaling, calibration with MDIA matching generally improved item parameter estimation of the operational items, regardless of the underlying θ distributions (as shown in the upper left section of the table). However, for new items, calibration with MDIA matching did not lead to more accurate estimation (but sometimes slightly worse estimation) in the pretest block (as shown in the upper right section of the table), which might have been caused by overfitting of the operational item responses. FIPC estimates with the reference sample ($N_0 = 5{,}000$) led to negligible error, at a magnitude of 0.03 to 0.04, for the operational items (as shown in the bottom left section of the table). FIPC estimates for the new items in the pretest block (as shown in the bottom right section of Table 1) outperformed other methods (calibration with SL scaling or with MDIA matching). Overall, estimation of new item parameters was most accurate when using FIPC with the original sample (without matching).

Table 1 Average MAEs When $J = 15$ and $N = 250$

	Operational items				Pretest items			
	a parameter		*b* parameter		*a* parameter		*b* parameter	
θ	SL(a)	M(a)	SL(b)	M(b)	SL(a)	M(a)	SL(b)	M(b)
PS	0.159	0.147	0.211	0.090	0.157	0.175	0.220	0.194
NS	0.221	0.130	0.182	0.126	0.221	0.225	0.198	0.207
$N(0, 1)$	0.191	0.133	0.184	0.104	0.181	0.190	0.197	0.202
$N\left(.5,1.2^2\right)$	0.190	0.127	0.184	0.098	0.185	0.194	0.186	0.195
$N\left(1,1.4^2\right)$	0.186	0.130	0.176	0.099	0.185	0.199	0.185	0.193
	F(a)	FM(a)	F(b)	FM(b)	F(a)	FM(a)	F(b)	FM(b)
PS	0.032	0.032	0.042	0.042	0.129	0.147	0.163	0.181
NS	0.032	0.032	0.044	0.044	0.124	0.146	0.171	0.187
$N(0, 1)$	0.032	0.032	0.044	0.044	0.125	0.139	0.172	0.189
$N\left(.5,1.2^2\right)$	0.031	0.031	0.041	0.041	0.119	0.138	0.165	0.185
$N\left(1,1.4^2\right)$	0.032	0.032	0.042	0.042	0.125	0.142	0.170	0.187

Note. MAE = mean absolute error, PS = positively skewed, NS = negatively skewed, SL = separate calibration with Stocking and Lord scaling, M = calibration with minimum discriminant information adjustment (MDIA) matching, F = fixed parameter calibration (FIPC), and FM = FIPC with matched data.

Table 2 Average MAEs When $J = 15$ and $N = 500$

	Operational items				Pretest items			
	a parameter		*b* parameter		*a* parameter		*b* parameter	
θ	SL(a)	M(a)	SL(b)	M(b)	SL(a)	M(a)	SL(b)	M(b)
PS	0.131	0.148	0.187	0.078	0.124	0.121	0.196	0.157
NS	0.189	0.109	0.143	0.123	0.191	0.173	0.155	0.168
$N(0, 1)$	0.148	0.114	0.140	0.089	0.142	0.140	0.149	0.149
$N\left(.5,1.2^2\right)$	0.149	0.125	0.145	0.093	0.143	0.145	0.151	0.149
$N\left(1,1.4^2\right)$	0.162	0.129	0.145	0.097	0.151	0.151	0.158	0.156

	Operational items	Pretest items			
		F(a)	FM(a)	F(b)	FM(b)
PS		0.098	0.106	0.129	0.132
NS		0.090	0.097	0.122	0.125
$N(0, 1)$		0.091	0.095	0.122	0.127
$N\left(.5,1.2^2\right)$		0.091	0.096	0.121	0.125
$N\left(1,1.4^2\right)$		0.088	0.093	0.120	0.126

Note. MAE = mean absolute error, PS = positively skewed, NS = negatively skewed, SL = separate calibration with Stocking and Lord scaling, M = calibration with minimum discriminant information adjustment (MDIA) matching, F = fixed parameter calibration (FIPC), and FM = FIPC with matched data.

Table 3 Average MAEs When $J = 15$ and $N = 1{,}000$

	Operational items				Pretest items			
	a parameter		*b* parameter		*a* parameter		*b* parameter	
θ	SL(a)	M(a)	SL(b)	M(b)	SL(a)	M(a)	SL(b)	MAE.M(b)
PS	0.100	0.131	0.181	0.071	0.088	0.086	0.191	0.138
NS	0.171	0.105	0.121	0.118	0.159	0.140	0.120	0.143
$N(0, 1)$	0.129	0.114	0.124	0.096	0.122	0.123	0.130	0.131
$N\left(.5,1.2^2\right)$	0.131	0.116	0.128	0.094	0.118	0.118	0.128	0.125
$N\left(1,1.4^2\right)$	0.134	0.116	0.121	0.090	0.118	0.117	0.121	0.120

	Operational items	Pretest items			
		F(a)	FM(a)	F(b)	FM(b)
PS		0.077	0.084	0.095	0.096
NS		0.060	0.063	0.085	0.085
$N(0, 1)$		0.066	0.067	0.091	0.091
$N\left(.5,1.2^2\right)$		0.062	0.064	0.085	0.087
$N\left(1,1.4^2\right)$		0.063	0.065	0.087	0.088

Note. MAE = mean absolute error, PS = positively skewed, NS = negatively skewed, SL = separate calibration with Stocking and Lord scaling, M = calibration with minimum discriminant information adjustment (MDIA) matching, F = fixed parameter calibration (FIPC), and FM = FIPC with matched data.

Tables 2 and 3 report the average MAEs for sample sizes of 500 and 1,000, respectively. Because calibration of the operational items with the reference sample stayed the same (as in the bottom left section of Table 1), this portion is not presented.

For the larger sample sizes, observations from Table 1 hold. That is, calibration with MDIA led to better item parameter estimation than calibration with SL scaling for the operational items, but not for the pretest new items. Among all the calibration methods, FIPC with the original sample performed the best across different underlying θ distributions.

Tables 1 to 3 also show that the item parameter estimation became more accurate as the sample size increased. As expected, as the sample size increased, the difference between calibration with SL scaling and calibration with MDIA

Table 4 FIPC Calibrations of Pretest Items When $J = 15$ and $N = 250$, 500, or 1,000

$N = 250$	MAE.1F(a)	MAE.2F(a)	MAE.1F(b)	MAE.2F(b)
NS	0.255	0.169	0.214	0.139
$N(0, 1)$	0.236	0.200	0.194	0.168
$N\left(.5,1.2^2\right)$	0.219	0.188	0.174	0.156
$N\left(1,1.4^2\right)$	0.231	0.199	0.195	0.154
$N = 500$	MAE.1F(a)	MAE.2F(a)	MAE.1F(b)	MAE.2F(b)
NS	0.167	0.133	0.151	0.112
$N(0, 1)$	0.142	0.130	0.125	0.106
$N\left(.5,1.2^2\right)$	0.132	0.129	0.119	0.109
$N\left(1,1.4^2\right)$	0.146	0.123	0.127	0.103
$N = 1{,}000$	MAE.1F(a)	MAE.2F(a)	MAE.1F(b)	MAE.2F(b)
NS	0.121	0.098	0.111	0.087
$N(0, 1)$	0.103	0.095	0.086	0.076
$N\left(.5,1.2^2\right)$	0.106	0.095	0.091	0.081
$N\left(1,1.4^2\right)$	0.104	0.095	0.087	0.075

Note. MAE = mean absolute error, 1F = fixed parameter calibration (FIPC) with data from one administration, 2F = FIPC with data from two administrations.

matching became smaller for both operational items and pretest new items, but FIPC outperformed both for new items. Moreover, item estimation from FIPC with 250 test takers was nearly as accurate as those from calibrations with SL scaling with 500 students. Similarly, item estimation from FIPC with 500 students was nearly as accurate as (sometimes more accurate than) those from calibrations with SL scaling with 1,000 students. Observations from average MAEs for tests with $J = 30$ in the routing block were largely similar to those for $J = 15$. Those results are presented in Table B1 in Appendix B.

Note that items in the target blocks were very unstable when sample sizes were small because of missing data by design, so they were not considered as anchor items for scaling nor fixed with item parameters in FIPCs. This choice not only simplified the item calibration procedure, but also met the requirement for item scaling; that is, the portion of anchor items was at least 20% of the total items in transformation (Kolen & Brennan, 2004).

Study 2: FIPC With Multiple Administrations

To improve the initial pretest item calibration in one administration, new items were recalibrated with data from two administrations. In the first FIPC (1F), the pretest items were administrated (Admin 1) to a sample of test takers (size$=N_1$) with Test Form 1, and the initial FIPC estimates were obtained by fixing the routing items on Test Form 1. In the second FIPC (2F), the pretest items were assembled into the target blocks in Test Form 2 according to their initial estimates and administered to a sample of test takers (size $= N_2$; refer to Case B in Figure 4). With data from the two administrations, a concurrent FIPC calibration (size$=N_1 + N_2$) was conducted for the new items, where item parameters in routing blocks of Test Form 1 and Test Form 2 were fixed (note that items in the routing blocks were assumed to be from the item pool with accurate estimates).

The average MAEs of item parameter estimates are shown in Table 4. From the top section of Table 4, we observed that, with additional data, the accuracy of item parameter estimation was much improved; that is, the MAEs were smaller in the combined data than those when the sample size was 250 for both administrations. Because of missing data by design in the target blocks in the second administration, the magnitude of improvement was not as large as that with doubled sample sizes. The same observation applies to sample sizes $N = 500$ and 1,000 in the middle and lower sections of Table 4.

Study 3: Conversion Table Comparison

To investigate whether it was appropriate to use the 1PL model to calibrate pretest items instead of the 2PL model, we compared the conversion tables produced by 1PL and 2PL IRT true score equating. Figure 6 shows the MAEs in the conversion tables (the y-axis) at each score point (the x-axis) when the sample size $N = 100$ was used in FIPCs, where the

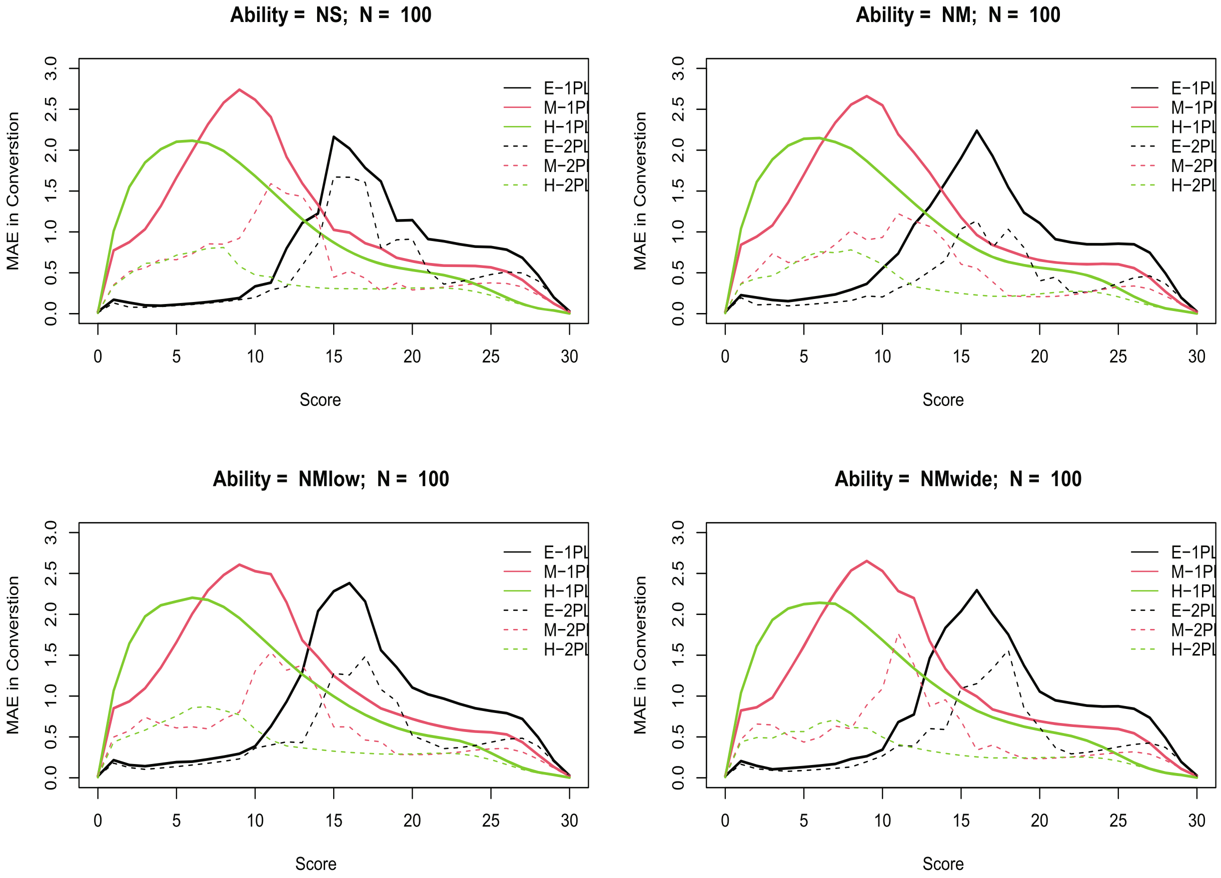

Figure 6 MAE in conversion tables ($N = 100$). *Note*. E-1PL, M-1PL, and H-1PL are produced by 1PL models on the easy, medium hard, and hard item blocks, respectively. E-2PL, M-2PL, and H-2PL are produced by 2PL models on the easy, medium hard, and hard item blocks, respectively. MAE = mean absolute error.

solid lines are produced by the 1PL model and the dashed lines by the 2PL model, for the forms with easy (black line), medium difficult (red line), and difficult (green line) target blocks. Note for Study 3, the conversion table produced by the true 2PL model item parameters is used as criterion for evaluation. From Figure 6, we observe that the 1PL model calibrations produced larger MAE than the 2PL calibration in the conversion tables when $N = 100$ from different ability distributions.

For larger sample sizes, the MAEs produced by the 2PL models became smaller and smaller, but the MAEs remained large for the 1PL models, as expected (refer to Figures 7 to 9).

Discussion and Conclusion

In the current research, we used item parameters from an MST program as illustration to evaluate various psychometric decision-making practices. We conducted three studies to investigate how to address the small-sample challenges for a specific testing program. In Study 1, we investigated three calibration methods under the MST design to estimate new items to replenish the item pool. Simulation results showed that the fixed item parameter calibration method (i.e., MWU-MEM) performed the best under different sample sizes and different underlying ability distributions, which agreed with results in several recent studies for linear tests (Kim & Kolen, 2019; König et al., 2021; Wang et al., 2020). More specifically, we found that the performance of FIPC was generally as accurate as that from separate calibrations with a doubled sample size for the studied conditions.

Therefore, when the item pool is accurately estimated with a 2PL model, as is the case for the studied program, FIPC is recommended for new item calibration in the MST design. In addition, we recommend using items in the routing block only as the fixed item parameters in practice because it leads to stable estimation, meets the required proportion of

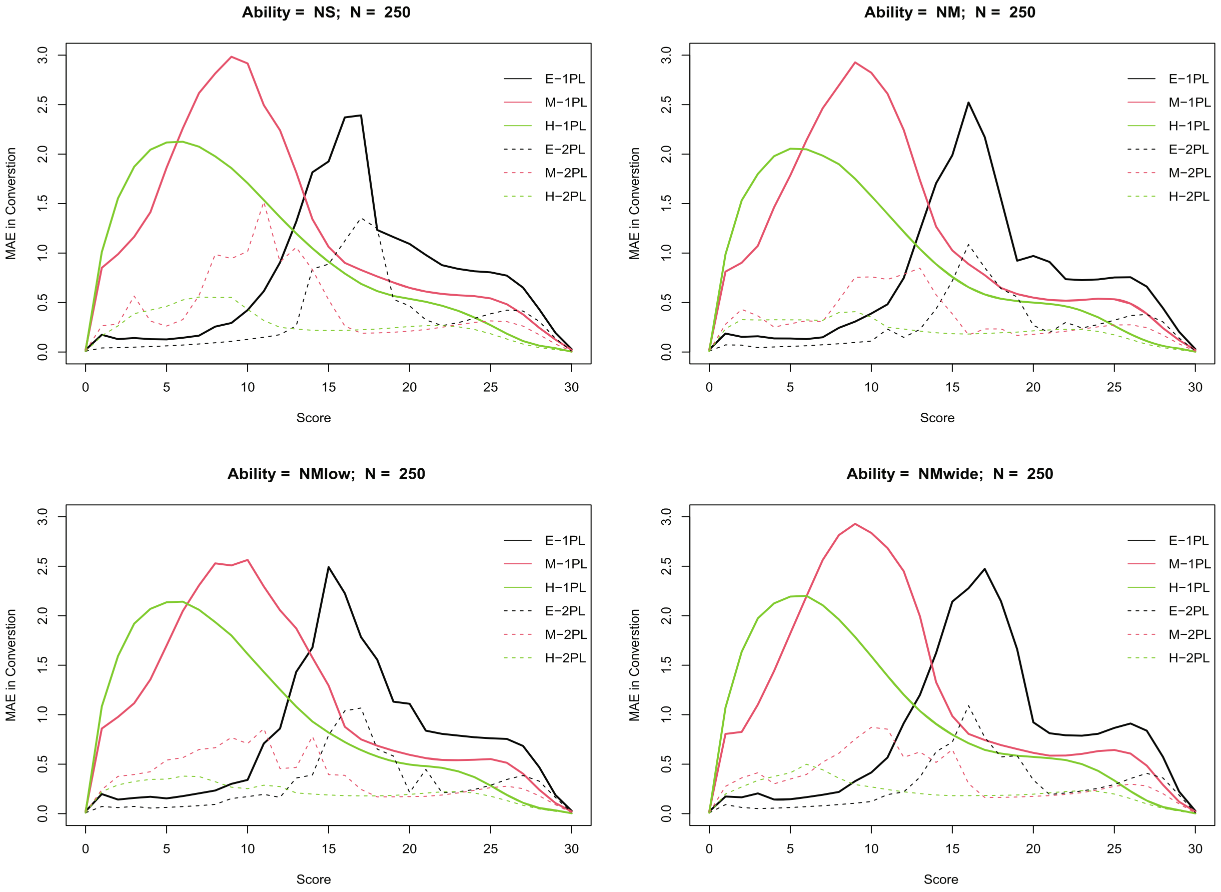

Figure 7 MAE in conversion tables ($N = 250$). *Note*. MAE = mean absolute error.

anchor items for item scaling (Kolen & Brennan, 2004), produces unbiased estimated parameters of new items under the MST design for meeting the missing at random requirement (Mislevy & Wu, 1988, 1996; Wang et al., 2020), and, most importantly, simplifies implementation of FIPC in practice. The results from Study 1 also showed that, if the program switched to the FIPC method, the sample size requirement (for example, $N = 1{,}000$) could be relaxed (say, $N > 500$), and the new item calibration could maintain quality similar to that when the separate calibration with SL was used. In case the initial new item calibration is not satisfying, Study 2 showed that concurrent FIPC calibration that combines data from multiple administrations can also improve the accuracy of new item parameter estimation.

The initial item pool of the studied program was well calibrated with field test data. Given small samples in operation in Study 3, we evaluated whether there was evidence to support a transition from the 2PL model to the 1PL model, in terms of the conversion table changes for score reporting. Based on the specific item pool, our limited simulation results showed that, unlike what was found in some previous literature, the 2PL model still performed better in terms of the conversion table accuracy. This finding might be attributed to the relatively large variation in the item discrimination parameters in the item pool, combined with the relatively short tests.

Based on our simulation results, one additional finding worth mentioning is that to post-calibrate operational items (i.e., existing items in the pool), it seems beneficial to use calibrations with MDIA matching, particularly when the sample size is small. In other words, if the goal of calibration is to recalibrate items in the pool, then the MDIA matching approach may be helpful. Furthermore, once items are all well calibrated from different administrations, the simultaneous linking method (Haberman, 2009) can be implemented to put the large number of item parameters in the item pool on the same scale to further improve the quality of the item pool.

One limitation for our findings in the current research is its generalizability to other MST testing programs. The studied item pools mostly contain easy items, even though we varied the ability distributions in the simulations, some findings may not apply to other program-specific data. However, the procedures for evaluating various psychometric decisions for

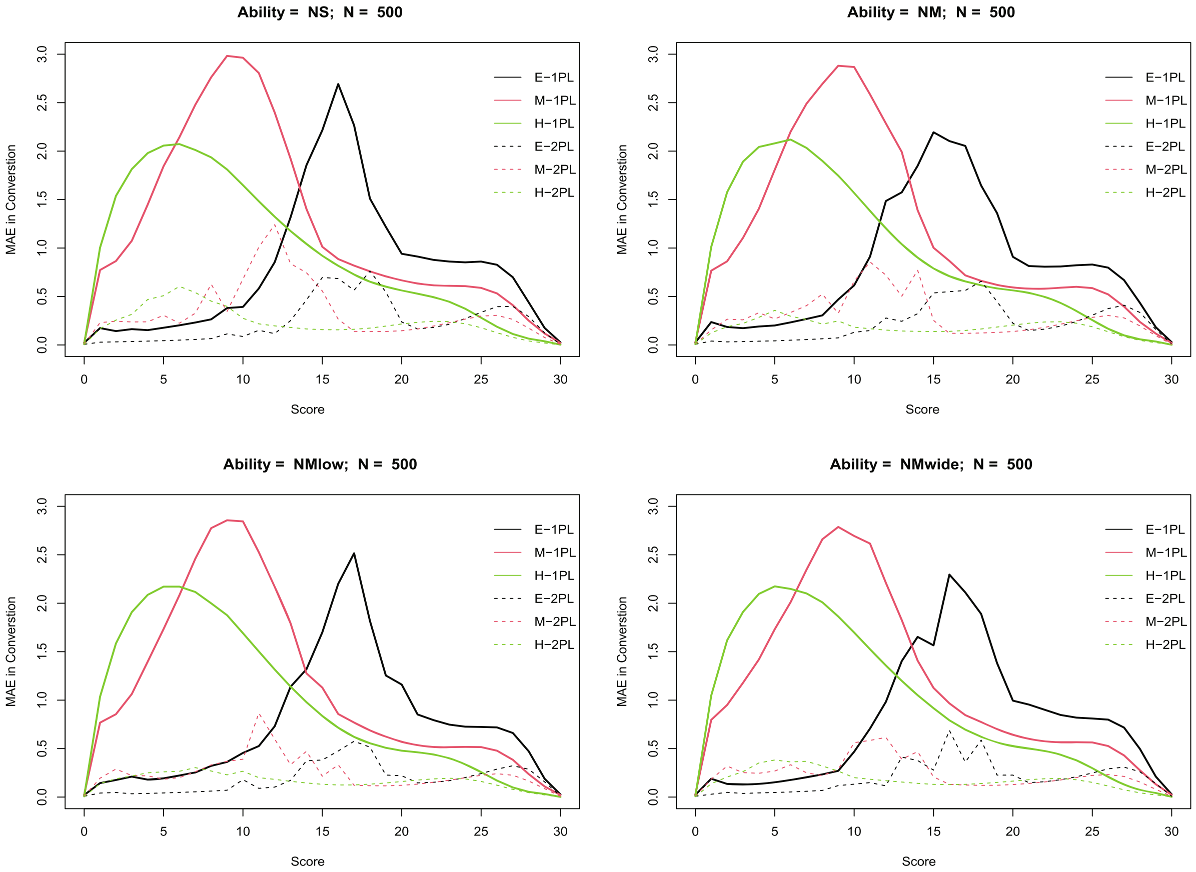

Figure 8 MAE in conversion tables ($N = 500$). *Note*. MAE = mean absolute error.

the specific program are applicable, and the recommendation of FIPC with the routing block is likely to be applicable as well for an MST program. The results of the model selection study (Study 3) may be different for different testing program settings.

Another limitation is the concurrent calibration design in Study 2 when using data collected from multiple administrations. In Study 2, we assumed that the new items first appeared in a pretest item block with Test Form 1 and then reappeared in the target blocks with Test Form 2 at the second administration, as depicted in Case B in Figure 4. This design would ensure accuracy of the routing decision for Test Form 2 and save the pretest block for newer items, and the recalibrated item parameters in the target blocks would also ensure accurate score reporting for Test Form 2. However, as we mentioned before, the gain of the Study 2 design (Case B in Figure 4) is at a price of smaller samples at the second stage for each target block because of missing data by design under MST. In addition, Study 2 with Case B design also requires the testing program to adjust operational procedures and timelines for score reporting because of the recalibration step. If the testing program can afford to readministrate the intact pretest block in Test Form 2, the sample size in the multiple-administration concurrent FIPC calibration would be doubled. If so, it is recommended to use Case A in Figure 4 in practice because of its simplicity. On the other hand, when a testing program has a shallow item pool and is in dire need of new items, the Case C design in Figure 4 for data collection might be considered because it can pretest the most item with each administration. As noted in Kim (2006), unstable parameter estimates of the fixed items due to small-sample sizes may not appear to have much effect on the performance of the FIPC methods in calibrating new items. Therefore, practitioners could embed some newly calibrated items with preliminary item parameters in the routing block with Test Form 2, to gather more data for a recalibration of these new items from Test Form 1, and save space in the pretest block on Test Form 2 to test newer items. However, impact of such a design on the routing decision accuracy needs to be further investigated under different scenarios with program-specific data.

As in most adaptive testing programs, the studied MST testing program uses the IRT true score equating in practice to link scores among different test forms. Therefore, in Study 3, we evaluated conversion table variation when different

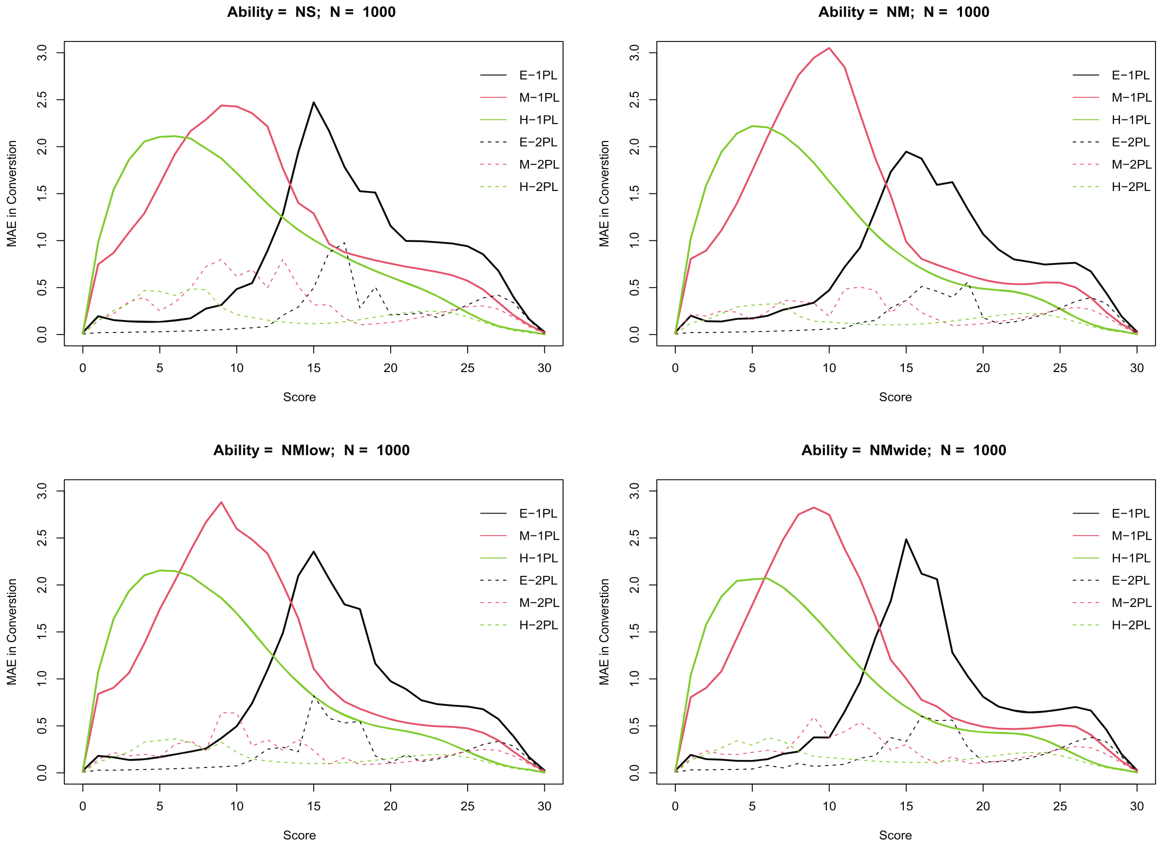

Figure 9 MAE in conversion tables (N = 1,000). *Note*. MAE = mean absolute error.

IRT models were used. The 2PL IRT true score equating method has an inherent issue when the test length is short (that is, plugging in the observed sum score, as if it were the true score, in the conversion table to obtain the scale score for reporting; Guo & Dorans, 2019, 2020; Kolen & Brennan, 2004; Lord, 1980). However, if a testing program uses the estimated ability or its derivatives for score reporting, further studies can use the program-specific data to investigate the benefit of using a simpler IRT model when the sample sizes are small.

Overall, when a testing program makes psychometric decisions or adjusts decisions when the testing situation changes, it is prudent to conduct program-specific and literature-guided research.

Acknowledgments

The authors would like to thank Shuhong Li, Ru Lu, and other colleagues for data-related discussions. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of ETS.

References

Baker, F. B., & Kim, S. H. (2004). *Item response theory: Parameter estimation techniques* (2nd ed.). CRC Press.

Ban, J.-C., Hanson, B. A., Wang, T., Yi, Q., & Harris, D. J. (2006). A comparative study of on-line pretest item-calibration/scaling methods in computerized adaptive testing. *Journal of Educational Measurement*, *38*(3), 191–212. https://doi.org/10.1111/j.1745-3984.2001.tb01123.x

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, *46*(4), 443–459. https://doi.org/10.1007/BF02293801

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, *48*(6), 1–29. https://doi.org/10.18637/jss.v048.i06

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*(1), 81–117. https://doi.org/10.1111/bmsp.12083

de Ayala, R. J. (2009). *T he theory and practice of item response theory*. Guilford Press.

DeMars, C. E., & Jurich, D. P. (2012). Software note: Using BILOG for fixed anchor item calibration. *Applied Psychological Measurement*, *36*(3), 232–236. https://doi.org/10.1177/0146621612438726

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. https://doi.org/10.1111/j.2517-6161.1977.tb01600.x

Drasgow, F. (1989). An evaluation of marginal maximum likelihood estimation for the two-parameter logistic model. *Applied Psychological Measurement*, *13*(1), 77–90. https://doi.org/10.1177/014662168901300108

Eggen, T. J. H. M., & Verhelst, N. D. (2011). Item calibration in incomplete testing designs. *Psicologica*, *32*(1), 107–132.

Ercikan, K., Guo, H., & He, Q. (2020). Use of response process data to inform group comparison and fairness research. *Educational Assessment*, *25*(3), 179–197. https://doi.org/10.1080/10627197.2020.1804353

Guo, H. (2022). How did students engage with a remote educational assessment? A case study. *Educational Measurement: Issues and Practice*, *41*(3), 58–68. https://doi.org/10.1111/emip.12476

Guo, H., & Dorans, N. (2019). *Observed scores as matching variables in differential item functioning under the one- and two-parameter logistic models: Population results* (Research Report RR-19-06). ETS. https://doi.org/10.1002/ets2.12243

Guo, H., & Dorans, N. (2020). Using weighted sum scores to close the gap between DIF practice and theory. *Journal of Educational Measurement*, *57*(4), 484–510. https://doi.org/10.1111/jedm.12258

Guo, H., & Ercikan, K. (2021). *Comparing test-taking behaviors of English language learners (ELLs) to non-ELL students: Use of response time in measurement comparability research* (Research Report No. RR-21-25). ETS. https://doi.org/10.1002/ets2.12340

Hambleton R. K., & Xing, D. (2006). Optimal and nonoptimal computer-based test designs for making pass-fail decisions. *Applied Measurement in Education*, *19*(3), 221–239. https://doi.org/10.1207/s15324818ame1903_4

Haberman, S. J. (1984). Adjustment by minimum discriminant information. *The Annals of Statistics*, *12*(3), 971–988. https://doi.org/10.1214/aos/1176346715

Haberman, S. J. (2009). *Linking parameter estimates derived from an item response model through separate calibrations* (Research Report No. RR-09-40). ETS. https://doi.org/10.1002/j.2333-8504.2009.tb02197.x

Haberman, S. J. (2015). Pseudo-equivalent groups and linking. *Journal of Educational and Behavioral Statistics*, *40*(3), 254–273. https://doi.org/10.3102/1076998615574772

Jewsbury, P. A., & van Rijn, P. W. (2020). IRT and MIRT models for item parameter estimation with multidimensional multistage tests. *Journal of Educational and Behavioral Statistics*, *45*(4), 383–402. https://doi.org/10.3102/1076998619881790

Jiao, H., & Lissitz, R. W. (2020). What hath the coronavirus brought to assessment? Unprecedented challenges in educational assessment in 2020 and years to come. *Educational Measurement: Issues and Practice*, *39*(3), 45–48. https://doi.org/10.1111/emip.12363

Kim, S. (2006). A comparative study of IRT fixed parameter calibration methods. *Journal of Educational Measurement*, *43*(4), 355–381. https://doi.org/10.1111/j.1745-3984.2006.00021.x

Kim, S., & Kolen, M. J. (2019). Application of IRT fixed parameter calibration to multiple-group test data. *Applied Measurement in Education*, *32*(4), 310–324. https://doi.org/10.1080/08957347.2019.1660344

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices.* Springer. https://doi.org/10.1007/978-1-4757-4310-4

König, C., Khorramdel, L., Yamamoto, K., & Frey, A. (2021). The benefits of fixed item parameter calibration for parameter accuracy in small sample situations in large-scale assessments. *Educational Measurement: Issues and Practice*, *40*(1), 17–21. https://doi.org/10.1111/emip.12381

Lord, F. M. (1980). *Applications of item response theory to practical testing problems. Erlbaum.* https://doi.org/10.4324/9780203056615

Lord, F. M. (1983). Small N justifies Rasch model. In D. Weiss (Ed.), *New horizons in testing: Latent trait test theory and computerized adaptive testing* (pp. 51–61). Academic Press. https://doi.org/10.1016/B978-0-12-742780-5.50011-1

Luecht R., Brumfield T., & Breithaupt K. (2006). A testlet assembly design for adaptive multistage tests. *Applied Measurement in Education*, *19*(3), 189–202. https://doi.org/10.1207/s15324818ame1903_2

Mead, A. D.(2006). An introduction to multistage testing. *Applied Measurement in Education*, *19*(3), 185–187. https://doi.org/10.1207/s15324818ame1903_1

Mislevy, R. J., & Bock, R. D. (1986). PC-BILOG: Item analysis and est coring with binary logistic models [Computer software]. Scientific Software.

Mislevy, R. J., & Sheehan, K. M. (1989). The role of collateral information about examinees in item parameter estimation. *Psychometrika* *54*(4), 661–679. https://doi.org/10.1007/BF02296402

Mislevy, R. J., & Wu, P.-K. (1988). *Inferring examinee ability when some item responses are missing* (Research Report No. RR-88-48-ONR). ETS. https://doi.org/10.1002/j.2330-8516.1988.tb00304.x

Mislevy, R. J., & Wu, P.-K. (1996). *Missing responses and IRT ability estimation: Omits, choice, time limits, and adaptive testing* (Research Report No. RR-96-30-ONR). ETS. https://doi.org/10.1002/j.2333-8504.1996.tb01708.x

Mullis, I. V. S., & Martin, M. O. (Eds.). (2019). *PIRLS 2021 assessment frameworks*. TIMSS & PIRLS International Study Center. https://timssandpirls.bc.edu/pirls2021/frameworks/

O'Neill, T. R., Gregg, J. L., & Peabody, M. R. (2020). Effect of sample size on common item equating using the dichotomous Rasch model. *Applied Measurement in Education*, *33*(1), 10–23. https://doi.org/10.1080/08957347.2019.1674309

Partchev, I., & Maris, G. (2017). *irtoys: A collection of functions related to item response theory (IRT)* (R package version 0.2.1) [Computer software]. https://CRAN.R-project.org/package=irtoys

Peabody, M. R. (2020). Some methods and evaluation for linking and equating with small samples. *Applied Measurement in Education*, *33*(1), 3–9. https://doi.org/10.1080/08957347.2019.1674304

Reise, S. P., Rodriguez, A., Spritzer, K. L., & Hays, R. D. (2018). Alternative approaches to ad- dressing non-normal distributions in the application of IRT models to personality measures. *Journal of Personality Assessment*, *100*(4), 363–374. https://doi.org/10.1080/00223891.2017.1381969

Seong, T. J. (1990). Sensitivity of marginal maximum likelihood estimation of item and ability parameters to the characteristics of the prior ability distributions. *Applied Psychological Measurement*, *14*(3), 299–311. https://doi.org/10.1177/014662169001400307

Stocking M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement*, *7*(2), 201–210. https://doi.org/10.1177/014662168300700208

Stone, C. A. (1992). Recovery of marginal maximum likelihood estimates in the two-parameter logistic response model: An evaluation of MULTILOG. *Applied Psychological Measurement*, *16*(1), 1–16. https://doi.org/10.1177/014662169201600101

von Davier, M., Sinharay, S., Oranje, A., & Beaton, A. (2006). The statistical procedures used in National Assessment of Educational Progress: Recent developments and future directions. In C. R. Rao & S. Sinharay (Eds.), Handbook of statistics: Psychometrics *(Vol. 26, pp. 1039–1055)*. Elsevier. https://doi.org/10.1016/S0169-7161(06)26032-2

Wainer, H., & Mislevy, R. J. (1990). Item response theory, item calibration, and proficiency estimation. In H. Wainer (Ed.), *Computerized adaptive testing: A primer* (pp. 65–102). Erlbaum.

Wang, C., Chen, P., & Jiang, S. (2020). Item calibration methods with multiple subscale multistage testing. *Journal of Educational Measurement*, *57*(1), 3–28. https://doi.org/10.1111/jedm.12241

Wise, S. L. (2021). Six insights regarding test-taking disengagement. *Educational Research and Evaluation*, *26*(5–6), 328–338. https://doi.org/10.1080/13803611.2021.1963942

Woodruff, D. J., & Hanson, B. A. (1996). *Estimation of item response models using the EM algorithm for finite mixtures* (Research Report No. 96-6). ACT.

Yamamoto, K., Shin, H. J., & Khorramdel, L. (2018). Multistage adaptive testing design in international large-scale assessments. *Educational Measurement: Issues and Practice*, *37*(4). 16–27. https://doi.org/10.1111/emip.12226

Appendix A

R Implementation

For implementation of the fixed parameter calibration (FIPC) in BILOG, PARSCALE, and ICL, please refer to Kim (2006), Kim and Kolen (2019), and references therein.

All analyses and calibrations in the current research were conducted by using the R language and its associated R packages. Therefore, R implementation of a few key steps is provided below.

In the current research, the *mirt* package (Chalmers, 2012) was used for item calibrations, and the *irtoys* package (Partchev & Maris, 2017) was used for the SL scaling. The MDIA method was programmed in R based on Haberman (2015).

Separate Calibration

The separate calibration use the *mirt* function with the default MMLE/EM method.

```
library(mirt)

#dat: new item responses on routing block and pretest blocks.
 mod.new<- mirt(dat, 1, verbose = FALSE)
 coef(mod.new) #estimated item parameters
```

FIPC

The *fixedCalib* function in *mirt* was used for FIPC.

```
#mod.ref: mirt output of the reference sample with the routing block items.
#dat: new item responses on routing block and pretest blocks.
MWU_MEM <- fixedCalib(dat, model = 1, old_mod = mod.ref)
coef(MWU_MEM) #estimated item parameters
```

SL Scaling

The *sca* function in *irtoys* was used for SL scaling.

```
library(irtoys)
#ITEM.ref: data matrix of the anchor item parameters
#(a, b, and c=0 are in first, second, and third column, respectively).
#ITEM.new: data matrix of item parameters
#(a, b and c=0 are in first, second, and third column, respectively).
J=length(ITEM.ref[1,]) #number of the anchor items
qq=normal.qu(n =61, lower = -6, upper = 6, mu = 0, sigma = 1,
                                     scaling = "points") #quadrature points
SL=sca(old.ip=ITEM.ref, new.ip=ITEM.new, old.items=1:J, new.items=1:J,
                                 old.qu=qq, method = "SL", bec=FALSE) #scaling
A<-SL$slope #slope coefficient
B<-SL$intercept #intercept
```

Appendix B

Results When $J = 30$

Table B1 Average MAEs for $J = 30$ in the Routing Block and $N = 250$ in Study 1

	Operational items				Pretest items			
θ	MAE.SL(a)	MAE.M(a)	MAE.SL(b)	MAE.M(b)	MAE.SL(a)	MAE.M(a)	MAE.SL(b)	MAE.M(b)
PS	0.133	0.163	0.177	0.170	0.135	0.179	0.182	0.219
NS	0.154	0.114	0.182	0.111	0.148	0.171	0.188	0.205
$N(0, 1)$	0.143	0.122	0.177	0.099	0.134	0.160	0.177	0.194
$N\left(.5,1.2^2\right)$	0.139	0.119	0.171	0.098	0.135	0.161	0.175	0.192
$N\left(1,1.4^2\right)$	0.143	0.119	0.171	0.097	0.136	0.163	0.179	0.196
	MAE.F(a)	MAE.FM(a)	MAE.F(b)	MAE.FM(b)	MAE.F(a)	MAE.FM(a)	MAE.F(b)	MAE.FM(b)
PS	0.034	0.034	0.046	0.046	0.129	0.157	0.168	0.202
NS	0.034	0.034	0.044	0.044	0.124	0.154	0.174	0.204
$N(0, 1)$	0.036	0.036	0.045	0.045	0.120	0.140	0.168	0.195
$N\left(.5,1.2^2\right)$	0.034	0.034	0.044	0.044	0.119	0.143	0.165	0.191
$N\left(1,1.4^2\right)$	0.037	0.037	0.045	0.045	0.121	0.145	0.172	0.197

Note. SL = separate calibration with Stocking and Lord scaling, M = calibration with minimum discriminant information adjustment (MDIA) matching, F = fixed parameter calibration (FIPC), and FM = FIPC with matched data.

Suggested citation:

Guo, H., Johnson, M. S., McCaffrey, D. F., & Gu, Lixong. (2024). *Practical considerations in item calibration with small samples under multistage test design: A case study* (Research Report No. RR-24-03). ETS. https://doi.org/10.1002/ets2.12376

Action Editor: Usama Ali

Reviewers: Ru Lu and Paul Jewsbury

Find other ETS-published reports by searching the ETS ReSEARCHER database.