A comparison of five methods for pretest item selection in online calibration

Yi Zheng*

Division of Educational Leadership and Innovation,
Mary Lou Fulton Teachers College,
Arizona State University,
1050 S. Forest Mall,
Tempe, AZ 85281, USA
Email: yi.isabel.zheng@asu.edu
*Corresponding author

Hua-Hua Chang

Department of Educational Psychology,
University of Illinois at Urbana-Champaign,
1310 S. 6th St,
Champaign, IL 61820, USA
Email: hhchang@illinois.edu

Abstract: Many long-term testing programs rely on large item banks that need to be replenished regularly with new items, and these new items need to be pretested before being used operationally. Online calibration is a pretesting strategy in computerised adaptive testing, which embeds pretest items in operational tests and adaptively matches the pretest items with examinees. This paper compares five existing methods for pretest item selection in online calibration. A simulation study was conducted under the one-, two-, and three-parameter logistic models. The effects of two estimation methods, three seeding locations, and five calibration sample sizes were also investigated. Findings from the simulation study are mixed. However, overall, the simplest random selection method appears to be a potential best choice.

Keywords: computerised adaptive testing; item parameter estimation; online calibration.

Reference to this paper should be made as follows: Zheng, Y. and Chang, H-H. (2017) 'A comparison of five methods for pretest item selection in online calibration', *Int. J. Quantitative Research in Education*, Vol. 4, Nos. 1/2, pp.133–158.

Biographical notes: Yi Zheng is currently an Assistant Professor at Arizona State University with joint appointment between the Mary Lou Fulton Teachers College and the School of Mathematical and Statistical Science. She received PhD in Educational Measurement and MS in Statistics from University of Illinois at Urbana-Champaign. She also served as a Managing Editor for Applied Psychological Measurement between 2011 and 2014. Her primary research area is educational testing designs, with specialty in computerised adaptive testing, multistage testing, item response theory, cognitive diagnostic testing and automated test assembly.

Hua-Hua Chang is a Professor in Educational Psychology, Psychology and Statistics at University of Illinois at Urbana-Champaign. His primary research area is educational measurement theories, computerised adaptive testing and cognitive diagnostic testing. Currently, he is also the Editor-in-chief of Applied Psychological Measurement.

This paper is a revised and expanded version of a paper entitled 'The ordered informative range priority index (OIRPI) method for item selection in online calibration', presented at *The Annual Meeting of the National Council on Measurement in Education*, Philadelphia, PA, 4–6 April 2014.

1 Introduction

As a long-term testing program continues, its item bank needs to be replenished regularly by replacing overexposed, obsolete, or flawed items with new items. When *item response theory* (IRT) is used to model test data, these new items need to be calibrated before being used operationally. In other words, their parameters need to be estimated and placed on the same scale as the operational items. Although it is possible to recruit examinees for the sole purpose of calibrating new items, a more cost-effective approach is to embed the new items in operational tests. This approach can also ensure that the examinees are almost equally motivated as in operational tests. Additionally, online calibration can be used if some of the operational items need to be recalibrated due to potential item parameter drift, which means their parameters may have changed through repeated administrations.

When applied in a *computerised adaptive test* (CAT) (Chang, 2004; Wainer, 2000), the aforementioned approach is specifically referred to as *online calibration* (Stocking, 1988). Analogous to the tailored testing feature in CAT where an optimal set of operational items is selected for each examinee to more efficiently estimate their ability levels, online calibration makes it possible to select an optimal sample of examinees for each pretest item to hopefully calibrate their parameter values more efficiently.

The exploration of online calibration methods started as early as almost three decades ago (Stocking, 1988), and researchers have mainly investigated two aspects of online calibration:

1 statistical methods for estimating item parameters in the online calibration setting

2 pretest item selection and seeding design.

The first aspect has been studied extensively (Ban et al., 2001, 2002; Chen et al., 2012; Chen and Wang, 2016; Segall, 2003; Stocking, 1988), whereas the second aspect is still being explored. To date several pretest item selection designs have been proposed (Chang and Lu, 2010; Chen et al., 2012; Kingsbury, 2009; Linden and Ren, 2015; Zheng, 2014), but there has not been a consensus regarding the optimal practices in various situations.

The purpose of this study is to compare five existing pretest item selection methods proposed in the recent literature through computer simulation under various settings.

The remaining of the paper is organised in the following way: first an overview of online calibration is provided. Then the five compared pretest item selection methods are reviewed, which include:

1 random selection

2 the 'examinee-centred' method

3 the original D-optimal method

4 the Bayesian D-optimal design (Linden and Ren, 2015)

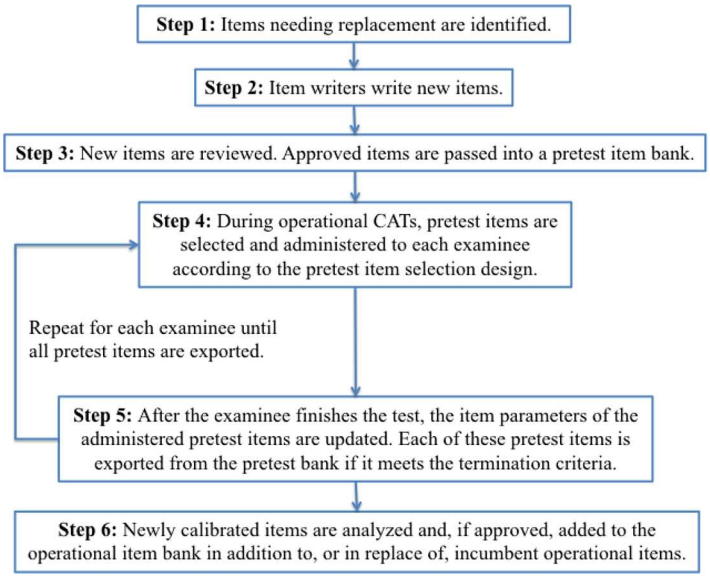
5 the ordered informative range priority index method (Zheng, 2014).

Finally, the simulation study that compares the five methods under one-, two-, and three-parameter logistic models, across two estimation methods, three seeding locations, and five calibration sample sizes are presented. Results reveal varied patterns under different treatment conditions. The findings are useful in not only the CAT context but also all the other computerised testing modes.

2 Overview of online calibration

The general procedure of online calibration for item bank replenishment is summarised by Figure 1. As the figure presents, Steps 1–3 pertain to preparing the pretest item bank. To recalibrate some operational items instead of replenishing the item bank, just skip Steps 2 and 3. Step 4 is the sampling step: during the operational test, when an examinee reaches a seeding location, pretest items are selected from the pretest item bank based on certain item selection rules. When the examinee finishes the test, Step 5 is carried out, where the item parameters of those administered pretest items are updated. Steps 4 and 5 are repeated for every new examinee, and the sampling outcome is constantly adjusted based on the updated parameter values. The sampling process of each pretest item can be terminated and the item be exported from the pretest item bank separately once a satisfactory accuracy of the parameter estimates is achieved or the maximum sample size is reached. As a final step (Step 6), the exported pretest items are reviewed and if approved, put into operational use.

Figure 1 The steps of item bank replenishment (see online version for colours)



Note that because any adaptive item selection rule adopted in Step 4 relies on provisional item parameter values, initial item parameters are needed at the first time. There are multiple ways to obtain initial parameter estimates. For example, item developers may provide educated guesses of the parameter values or an initial random selection phase can be implemented, at the end of which item parameters are estimated and to be taken to the next adaptive selection phase as the initial parameter values (Ban et al., 2001; Chen et al., 2012; Kingsbury, 2009).

Also note that in real applications of large-scale testing programs, it is possible that several examinees are taking the test almost simultaneously. This scenario is compatible with the online calibration workflow above because the calibration and update of the pretest item parameters do not have to be carried out after each individual examinee finishes his/her test. The new response data can be simply recorded, and calibration can be carried out after a batch of new data is obtained. In fact, this batch-based workflow is expected to be preferred for its ease on server computation load.

The aforementioned process belongs to the family of *sequential optimal sampling design* (Berger, 1992; Buyske, 2005; Jones and Jin, 1994). Similar methods have also been used in medical trials for decades to save research cost and time (Armitage, 1975). Given limited resources (e.g. examinees, time), this sequential optimal sampling design could potentially increase the accuracy of the calibrated item parameters. In other words, to achieve the same calibration accuracy, fewer examinees may be required than that is required in simple random sampling, the latter being typical in traditional paper-and-pencil non-adaptive tests. Moreover, by assigning different pretest items to each individual, this adaptive online calibration should pose less test security risk than assigning the same block of pretest items to a convenient sample (e.g. a school, a district) as often done in paper-and-pencil tests.

3 Review of existing pretest item selection methods

The existing pretest item selection methods in online calibration can be summarised into three categories:

1 random selection

2 examinee-centred adaptive selection

3 item-centred adaptive selection.

3.1 Random selection method

With random item selection (Ban et al., 2001), pretest items are randomly selected when a examinee reaches the seeding locations in the test. This method is the easiest to carry out, and when the sample size is large enough, it renders a calibration sample that follows a similar distribution as the examinee population.

3.2 Examinee-centred adaptive selection methods

With an examinee-centred adaptive selection method, pretest items are selected by the same item selection method as used for selecting operational items (Chen et al., 2012; Kingsbury, 2009). The operational item selection criteria in CAT are designed to optimise the estimation of examinee abilities (therefore this method is called 'examinee-centred' here), but not for the purpose of calibrating pretest items.

Conceptually, the examinee-centred method should be a reasonable choice for the one-parameter logistic (1PL) model. The item response function for the 1PL model is

$$P_j(\theta_i) = \frac{1}{1 + \exp\left[-(\theta_i - b_j)\right]}\,, \tag{1}$$

where $P_j(\theta_i)$ denotes the probability of examinee i correctly responding to item j, θ_i denotes the ability level of examinee i, and b_j denotes the difficulty of item j. A typical 1PL CAT optimises the estimation efficiency of examinee ability estimation by matching the item difficulty b with the estimated examinee's ability level $\hat{\theta}$. It is easy to see from the model equation that using the same method for selecting pretest items will also optimise the estimation efficiency of item parameter calibration.

However, in other IRT models, the θ values that constitute the optimal samples are different for each item parameter. For example, a three-parameter logistic (3PL) model (Eq. 2) item j has three parameters: the discrimination parameter (a_j), the difficulty parameter (b_j) and the pseudo-guessing parameter (c_j).

$$P_j(\theta_i) = c_j + \frac{1 - c_j}{1 + \exp\left[-a_j(\theta_i - b_j)\right]}\,. \tag{2}$$

The Fisher information for estimating the three item parameters, I_{aaij} for the a-parameter, I_{bbij} for the b-parameter and I_{ccij} for the c-parameter, is given below (Hambleton et al., 1991).

$$I_{aaij} = -E\left[\frac{\partial^2 \ell_{ij}}{\partial a_j \partial a_j}\right] = (\theta_i - b_j)^2 \frac{1 - P_j(\theta_i)}{P_j(\theta_i)} \left[\frac{P_j(\theta_i) - c_j}{1 - c_j}\right]^2\,, \tag{3}$$

$$I_{bbij} = -E\left[\frac{\partial^2 \ell_{ij}}{\partial b_j \partial b_j}\right] = a_j^2 \frac{1 - P_j(\theta_i)}{P_j(\theta_i)} \left[\frac{P_j(\theta_i) - c_j}{1 - c_j}\right]^2\,, \tag{4}$$

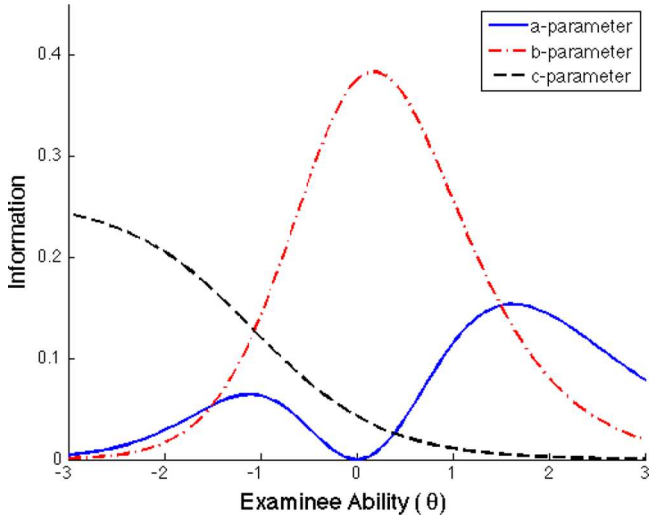
$$I_{ccij} = -E\left[\frac{\partial^2 \ell_{ij}}{\partial c_j \partial c_j}\right] = \frac{1 - P_j(\theta_i)}{P_j(\theta_i)} \frac{1}{(1 - c_j)^2}\,, \tag{5}$$

where $P_j(\theta_i)$ is given by Eq. (2), and ℓ_{ij} denotes the log-likelihood of observing a response to item j from examinee i with ability levels θ_i.

Figure 2 illustrates the information function for a 3PL model item with a set of reasonable item parameter values: $a = 1.5$, $b = 0$, $c = 0.2$. Fisher information is a measure of the discriminating power the observed random variables carry to distinguish the true value from nearby values of an unknown parameter. The higher the information, the more useful the corresponding ability θ is in estimating the item parameter. As Figure 2 shows, the peaks of the Fisher information curves typically occur at different θ locations for the

three parameters (also see Stocking, 1990) of the 3PL model. However, when operational items are selected in a 3PL CAT, the commonly adopted maximum Fisher information method essentially leads to a practical result of roughly matching b with $\hat{\theta}$ (Chang and Ying, 2008). This will provide a good amount of information for b but little information for either a or c. Thus for such IRT models, a pretest item should instead be assigned to a group of examinees, so that they will provide enough information in estimating all item parameters, ideally achieving a balanced maximisation according to certain statistical criterion. Selecting an inappropriate group of examinees could lead to inefficient or even seriously inaccurate item parameter estimation.

Figure 2 The information curves for a 3PL model item ($a = 1.5, b = 0, c = 0.2$) (see online version for colours)



Note: The curve for the c-parameter is scaled down by 10 times for better presentation.

3.3 Item-centred adaptive selection methods

In contrast to examinee-centred selection methods, item-centred adaptive selection methods select pretest items based on criteria directly designed to optimise the estimation of the pretest item parameters. So far the so-called *D-optimal* criterion has been the most frequently adopted in both general optimal calibration design literature (Berger, 1992; Berger et al., 2000) and in online calibration literature (Chang and Lu, 2010; Jones and Jin, 1994; Zhu, 2006). Using this criterion, different practical procedures have been proposed for the online calibration setting. The first section below introduces the D-optimal criterion, and the subsequent sections introduce the existing practical procedures that rely on the D-optimal criterion and thus belong to the group of item-centred selection methods.

3.3.1 D-optimal criterion

The D-optimal criterion is a traditional criterion in the subject of *optimal design* (Silvey, 1980). The idea of the D-optimal criterion is minimising the *generalised variance* (i.e. volume of the confidence ellipsoid) Anderson (1984) of parameter estimates by

maximising the determinant of the *Fisher information matrix*. In the IRT online calibration context, the D-optimal criterion is the determinant of the Fisher information matrix of the item-parameter vector given the ability parameter θs of all currently sampled examinees. Specifically, the information matrix of the item parameter vector (a_j, b_j, c_j) of a 3PL model item j provided by θ_i is given by the following equations (Hambleton et al., 1991):

$$\boldsymbol{I}_j(\theta_i) = \begin{pmatrix} I_{aaij} & I_{abij} & I_{acij} \\ I_{abij} & I_{bbij} & I_{bcij} \\ I_{acij} & I_{bcij} & I_{ccij} \end{pmatrix} . \tag{6}$$

While I_{aaij}, I_{bbij} and I_{ccij} have been given by Eqs. (3)–(5), the other components in the matrix are given below.

$$I_{abij} = -E\left[\frac{\partial^2 \ell_{ij}}{\partial a_j \partial b_j}\right] = -a_j\,(\theta_i - b_j)\,\frac{1 - P_j(\theta_i)}{P_j(\theta_i)}\left[\frac{P_j(\theta_i) - c_j}{1 - c_j}\right]^2 , \tag{7}$$

$$I_{acij} = -E\left[\frac{\partial^2 \ell_{ij}}{\partial a_j \partial c_j}\right] = (\theta_i - b_j)\,\frac{1 - P_j(\theta_i)}{P_j(\theta_i)}\,\frac{P_j(\theta_i) - c_j}{(1 - c_j)^2} , \tag{8}$$

$$I_{bcij} = -E\left[\frac{\partial^2 \ell_{ij}}{\partial b_j \partial c_j}\right] = -a_j\,\frac{1 - P_j(\theta_i)}{P_j(\theta_i)}\,\frac{P_j(\theta_i) - c_j}{(1 - c_j)^2} , \tag{9}$$

where ℓ_{ij} denotes the log-likelihood of observing a response to item j from examinee i with ability levels θ_i, and $P_j(\theta_i)$ denotes the item response function defined in Eq. (2). The corresponding formulas for the 1PL model and the 2PL model can be reduced from the formulas above.

As mentioned earlier in this paper, based on the assumption that the responses to an item from different examinees are independent, the Fisher information matrix of item j parameters given a vector of θs from N examinees (denoted by $\boldsymbol{I}$ below) is the direct summation of the Fisher information matrices of item j given each individual θ_i. Thus the D-optimal criterion is the determinant of the summed matrix:

$$D_j = |\boldsymbol{I}| = \left|\sum_{i=1}^{N} \boldsymbol{I}_j(\theta_i)\right| . \tag{10}$$

D_j provides a scalar summary of the information for estimating the parameters of item j. A greater D_j value indicates higher information, which is associated with a smaller generalised variance of the vector of item parameter estimates, defined as $|\mathrm{Cov}(\hat{\boldsymbol{\beta}})|$ (Anderson, 1984), where $\hat{\boldsymbol{\beta}}$ denotes the estimated item parameter vector:

$$\underset{\theta}{\operatorname{argmax}}\,|\boldsymbol{I}| = \underset{\theta}{\operatorname{argmin}}\,\frac{1}{|\boldsymbol{I}|} = \underset{\theta}{\operatorname{argmin}}\,|\boldsymbol{I}^{-1}| \xrightarrow{asy} \underset{\theta}{\operatorname{argmin}}\,|\mathrm{Cov}(\hat{\boldsymbol{\beta}})| . \tag{11}$$

In other words, the higher the D-optimal criterion value is, the smaller the estimation error for item parameter calibration is. Holding the sample size constant, a higher D-optimal criterion value should lead to a more efficient calibration.

Note that Eq. (11) holds when $\boldsymbol{I}$ is computed using the true values of examinee ability θ's and item parameter vector $\boldsymbol{\beta}$. In online calibration, however, true parameter values are

never available, and a sequential procedure as described in Figure 1 is adopted to gradually approximate the real optimal design - when computing $\boldsymbol{I}$, the θ values estimated from operational items are used in place of the true θ values, and the provisional estimates of item parameters are used in place of the true item parameter values. Fortunately, Ying and Wu (1997) have shown that under regularity conditions, this sequential design converges to the real optimal design as provisional item parameter estimates are updated more data accumulate. In addition, Chang (2011) further proved that under regularity conditions, the sequential design is asymptotically consistent and efficient when measurement errors of θ are present, that is, when estimated θ values are used in place of true θ values. These findings provide the mathematical foundation for the proposed method. For more details, readers can also refer to Berger (1991).

3.3.2 Early literature: online calibration as a sampling design

Some early online calibration literature (Chang and Lu, 2010; Jones and Jin, 1994; Zhu, 2006) treated online calibration as a sampling design, where the goal was to identify the optimal θ points (termed 'design space' in optimal design literature) for calibrating each item. Their designs are different from what is described in Section 2. In their design, each pretest item is handled separately, and for an item j, each sample point is selected sequentially where the kth sample (i.e. θ_k) is selected to maximise a certain measure of information. For those that used the D-optimal criterion, the quantity to be maximised is the following:

$$\Big|\sum_{i=1}^{k-1}\boldsymbol{I}_j(\theta_i)+\boldsymbol{I}_j(\theta_k)\Big|, \tag{12}$$

where the first component is the summation of information matrices provided by the $k-1$ existing samples item j has obtained and the second component is the information matrix provided by the possible new design point θ_k in the design space. Note that based on the assumption that the responses to an item from different examinees are independent, Fisher information has the following additive property: the Fisher information matrix of item j parameters given a vector of θs from N examinees is the direct summation of the Fisher information matrices of item j given each individual θ_i.

Deriving from the equation above, a few other computationally simpler designs were proposed additionally (Berger, 1992; Zhu, 2006). However, these early designs are in fact hardly feasible in the online calibration scenario. This is because all these designs assume there is an 'examinee pool' filled with examinees at various ability levels; for each pretest item, *the examinees in the examinee pool are compared* and the ones whose ability levels optimise the chosen criterion are selected. However, note that in practice, most CATs are administered continuously at scattered times. There can rarely be a static examinee pool to choose examinees from. Therefore, these designs are hardly feasible beyond simulation.

3.3.3 van der Linden and Ren's (2015) Bayesian D-optimal design

Based on the D-optimal idea, (Linden and Ren, 2015) proposed a procedure that is practically feasible. Their design follows the flow of continuous administration of an operational CAT and consequently the online calibration workflow described in Section 2. Examinees take the test at different times and reach seeding locations successively.

Whenever an examinee reaches a seeding location, *all pretest items are compared* and the item that maximises the so-called Bayesian D-optimal statistic value is selected. Specifically, the following Bayesian D-optimal statistic is computed for each pretest item j:

$$\Big| \sum_{i=1}^{k_j-1} \boldsymbol{I}_j(\hat{\theta}_i) + \boldsymbol{I}_j(\hat{\theta}) \Big| - \Big| \sum_{i=1}^{k_j-1} \boldsymbol{I}_j(\hat{\theta}_i) \Big| . \tag{13}$$

where $\hat{\theta}$ denotes the ability estimate of the current examinee who reaches a seeding location, and $\hat{\theta}_i$'s denote the ability estimates of all the $k_j - 1$ examinees who previously took item j. $k_j - 1$ does not have to be the same for all J pretest items - at a certain time point, each pretest item may have accumulated a different number of samples.

Among the pretest items, some items tend to render consistently higher D-optimal statistic values than others, caused by their superior estimated parameter values. Consequently, this design tends to select those items and ignore others. If a test developer terminates the calibration phase, some items in the pretest item pool could have very good parameter estimates but others may have no parameter estimates or extremely unreliable parameter estimates.

3.3.4 The ordered informative range priority index

To address some of the issues present in the above-mentioned methods, the ordered informative range priority index (OIRPI) method was developed from a new perspective - a 'need-based' perspective (Zheng, 2014). The OIRPI method selects the pretest item that 'needs' the incumbent examinees most.

Specifically, the OIRPI method follows the general workflow described in Section 2: when an examinee reaches a seeding location, an OIRPI index is computed for each pretest item in the pretest pool, and the item with the largest OIRPI value is selected for the incumbent examinee. The OIRPI value quantifies *how badly an item needs the incumbent examinee* by how informative this examinee is to this item *compared to potential examinees at other ability levels*. For example, if the θ value of the incumbent examinee is expected to produce higher information for item j than other θ values, item j has a high demand for this examinee because if it misses this examinee it is not very likely to receive future examinees who can provide such high information.

The following paragraphs describe the steps of the OIRPI method with the D-optimal criterion as the indicator of information. But in fact OIRPI is a framework that can be combined with any other appropriate criterion.

Step 1: The first step of the OIRPI method is to divide the examinee ability scale θ into several contiguous ranges and determine the representative θ value in each range. One way to divide the θ scale is equal spacing by the θ value and the representative values are the middle points of each range. Another way is equal spacing by the percentiles, for example, if 10 ranges are to be created, let P_t denote the tth percentile, the 10 ranges are $P_0 \sim P_{10}$, $P_{10} \sim P_{20}$, $\cdots$, $P_{90} \sim P_{100}$, and the representative values are $P_5, P_{15}, \cdots, P_{95}$. These percentiles can be obtained through empirical distribution of θ values from previous test data, or an assumed distribution such as normal distribution.

Then, when an examinee reaches a seeding location, Steps 2.1 through 2.3 are carried out **for each item** j to obtain their OIRPI values.

Step 2.1: Calculate D_{jr} given below provided by each θ-range r using their representative θ values θ_r.

$$D_{jr} = \Big| \sum_{i=1}^{k_j-1} \boldsymbol{I}_j(\hat{\theta}_i) + \boldsymbol{I}_j(\theta_r) \Big|, \tag{14}$$

where $\hat{\theta}_i$'s are the ability parameter estimates of the $k_j - 1$ samples item j has already accumulated, and the information matrices are calculated by the provisional item parameter estimates.

Step 2.2: Scale the information for all θ-ranges by $S_{jr} = (D_{jr} - min(D_{jr}))/(max(D_{jr}) - min(D_{jr}))$.

Step 2.3: Identify which range the estimated ability level of the incumbent examinee belongs to. Then assign the scaled information value of that range as the priority index of this item.

Step 3: After the OIRPI value is calculated for all pretest items, the item with the highest OIRPI value is selected for the incumbent examinee.

4 Simulation study

4.1 Design

A simulation study was conducted under the 1PL, 2PL and 3PL models using a Fortran program written by the first author to compare five pretest item selection methods:

1 random selection ('Random')

2 the examinee-centred method ('Examinee')

3 the original D-optimal design ('D-optima', as given in Eq. (12))

4 van der Linden and Ren's (2015) Bayesian D-optimal design ('B-D-optimal', as given in Eq. (13))

5 the OIRPI method ('OIRPI').

The third through fifth methods have been described in the previous section. Regarding the examinee-centred method, because the classical maximum Fisher information method (i.e. selecting the item that provides the maximum Fisher information for estimating examinee ability θ) was used as the operational item selection method in this study, the examinee-centred method used the same method.

The simulation study also included three other factors. The second factor is the method for estimating the pretest item parameters. This study included two most popular methods: the *one EM cycle method* (OEM) (Wainer and Mislevy, 2000) and the *multiple EM cycle method* (MEM) (Ban et al., 2001). These algorithms utilise the existing parameter values of the operational items taken by each examinee and naturally put the calibrated parameter

values of the pretest items on the existing IRT scale, without the need for the linking procedures.

Specifically, each time an item is calibrated, the MEM method includes iterations of the E-steps and the M-steps. The *E-step* finds the marginal log-likelihood of the item parameters using the posterior θ distribution from every examinee who took this item. The *M-step* finds the item parameter vector (e.g. (a, b, c)) that maximises the posterior marginal log-likelihood. In the first EM cycle, the posterior ability distribution of each examinee is obtained from only the administered operational items. In subsequent cycles, the posterior ability distribution is obtained from both the operational items and the pretest item being calibrated. The E-step and M-step iterate until the algorithm converges. Here, convergence is defined as the largest absolute change in all parameters no greater than a small critical value (0.001 in this study). Note that the MEM method is similar with the *fixed parameter* calibration (Kim, 2006). The OEM method is essentially the first EM cycle of the MEM method, with no subsequent EM cycles.

There are two differences between the estimation methods implemented in this study and that in Ban et al. (2001). First, Ban et al. (2001) used random item selection and did not update item parameters sequentially, and therefore they estimated all pretest items together after all response data were collected; whereas in this study, each pretest item was estimated individually once it received 10 new responses. Second, (Ban et al., 2001) used certain fixed Bayesian priors for the item parameters; in this study, to mimic a real situation where the Bayesian priors of the item parameters may not be known, the parameters for the Bayesian prior were obtained by fitting the lognormal, normal and beta distributions to the *operational* a-, b- and c-parameters, respectively. This solution may be more informative than using the prior distributions with an arbitrary choice prior parameter values, and less subject to the self-validating problem in finding 'empirical priors' from the items being calibrated, which are the *pretest* items.

The third factor is seeding location. Three levels were chosen for seeding location with test length being 40 items:

a early in the test (items 6 through 10)

b in the middle of the test (items 19 through 23)

c late in the test (items 32 through 36).

For each examinee, five different pretest items were seeded in one of those seeding ranges. In real practice, a seeding strategy with more randomness may be more acceptable than this fixed seeding because the latter may lead to differentiated motivation if the seeding locations are known by examinees. However, the seeding locations were fixed in this simulation study to better reveal the effect of seeding locations. These three conditions are expected to generate different results because the θ estimates, which are used in selecting pretest items, are of different levels of accuracies at different stages of the test. Also note that regardless of the seeding location, the pretest items are calibrated when an examinee finishes the entire test, and therefore all response data, including operational items that are administered after the seeded pretest items, are used in the calibration step.

To better distill patterns at different a-parameter and b-parameter values, the true item parameters of the pretest items were fixed at a set of discrete values. Specifically, for the 3PL model, a-parameters were set at four levels (0.5, 1.0, 1.5, 2.0), b-parameters were set

at nine levels (-2.0, -1.5, -1.0, -0.5, 0, 0.5, 1.0, 1.5, 2.0), and c-parameters were set constant at 0.2. Consequently, 36 pretest items were generated by fully crossing the levels of the three parameters. Similarly, 36 pretest items were generated in the same way for the 2PL model, with the c parameters set to 0. To keep the conditions consistent across the IRT models, 36 pretest items were also generated for the 1PL model with the same b-parameters as the other two models, a-parameters set to 1, and c-parameters set to 0.

The fourth factor is calibration sample size. In this simulation study, item parameter estimates were recorded at five different time points: when 1000, 1500, 2500, 5000, 7500 examinees have taken the test. No termination rule either based on sample size or measurement accuracy was imposed in the simulation in order to elicit purer effects of the factors. In online calibration, all pretest items will accumulate response samples as more examinees take the test successively, but they may accumulate samples at different rates. Nevertheless, given that each examinee takes five pretest items and there are a total of 36 pretest items, the chosen conditions correspond to an average of about 140, 200, 350, 700, 1000 calibration samples per pretest item.

Because parameter estimation is highly unstable when the sample size is too small, for the first 720 examinees, which means on average the first $720 \times 5/36 = 100$ responses each pretest item receives, pretest items were selected randomly and the item parameters were not updated until the end of this initial phase. After that, different adaptive pretest item selection methods were used, and the item parameters of each pretest item were updated after it obtained every 10 new samples. The pretest item selection criterion values were always computed using the latest parameter values.

The simulation was replicated for 100 times. In each replication, 300 operational items were randomly generated from the following distributions. These distributions were chosen to mimic realistic situations based on the empirical multivariate distributions found in the calibrated item parameters of a retired item bank as well as suggested by previous studies (Chang et al, 2001; Linden and Glas, 2000; Wingersky and Lord, 1984).

$$\begin{bmatrix} \log(a) \\ b \end{bmatrix} \sim MVN\left(\begin{bmatrix} 0.4 \\ 0.0 \end{bmatrix}, \begin{bmatrix} 0.10 & 0.15 \\ 0.15 & 1.00 \end{bmatrix} \right), \tag{15}$$

and

$$c \sim \text{Beta}(4, 16), \tag{16}$$

where the c-parameters are independent from the a- and b-parameters.

In each replication, examinee ability θs were also regenerated from the standard normal distribution. In the simulation, examinees took the test sequentially. During the test, operational items were selected from the operational item bank and the examinee's ability parameter was updated after each operational item was administered. *Maximum likelihood estimation* (MLE) (Baker and Kim, 2004) was the main method used to estimate θ. In addition, to stabilise the estimates when few data are available and to avoid MLE's difficulty with monotone likelihood, *expected a-posteriori* (EAP) (Baker and Kim, 2004) was used in place of MLE when the number of administered operational items was no more than five or when responses were all correct or all incorrect.

4.2 Evaluation criteria

Simulation results were analysed for each condition crossed by item selection method, seeding location, calibration sample size, and true a- and b-parameter values. Results were evaluated through two criteria. The first criterion focuses on the accuracy of the individual item parameter estimates. Specifically, root mean squared errors (RMSEs) of each item parameter's estimates, formulated by Eq. (17), are evaluated.

$$RMSE_p = \sqrt{\frac{1}{100}\sum_{r=1}^{100}\frac{1}{J}\sum_{j=1}^{J}(\hat{\beta}_{jrp} - \beta_{jrp})^2}\,, \tag{17}$$

where p denotes the specific element in the item parameter vector, such as the a-parameter, b-parameter and c-parameter; r denotes the replications; j denotes the pretest items with certain true a- and b-parameter values. For example, for the 2PL or 3PL models, there was only one pretest item in each replication for any combination of true a- and b-parameter values; for the 1PL model, there were 4 pretest items sharing the same true b-parameter value in each replication.

The second criterion focuses on the overall recovery of item parameter vectors. The chosen criterion is the average weighted area difference between the true *item characteristic curve* (ICC) and the estimated ICC. The area difference is computed by numeric integration using the following formula:

$$\text{ICCD} = \frac{1}{100}\sum_{r=1}^{100}\frac{1}{J}\sum_{j=1}^{J}\sum_{q=1}^{Q}|\hat{P}_{jr}(\theta_q) - P_{jr}(\theta_q)|g(\theta_q)\,, \tag{18}$$

where r denotes the replications, j denotes the pretest items with certain true a- and b-parameter values, q denotes 601 equally spaced quadrature points from -3 to 3, and the weighting function $g(\theta_q)$ is the density of the standard normal distribution. This weighting strategy will reflect the overall effect of the item parameter recovery on the entire normally distributed population. For the 2PL and 3PL models, ICC area differences will be used as the single-number summary to compare the performance of pretest item selection methods.

5 Results

The effects of the estimation method, seeding location, and calibration sample size are analysed first, so that a more focused analysis of the pretest item selection methods can be rendered afterwards.

5.1 Comparing estimation methods

The RMSEs of the two estimation methods (i.e. OEM and MEM) were investigated for all combinations crossed by other factors (i.e. true a- and b- parameter values, sample sizes, seeding locations, and pretest item selection methods). In most conditions, OEM generated a similar level of accuracy as MEM. The bigger discrepancies are seen in the RMSEs of the a-parameters for the 2PL and 3PL models. When the true a-parameter was small (i.e. $a = 0.5, 1.0$), OEM generated slightly more accurate parameter estimates than

MEM (RMSEs of OEM being about 0.05 to 0.1 lower than those of MEM). When the true a-parameter was large (i.e. $a = 2.0$), this difference is reversed and MEM was more accurate (RMSEs of MEM being about 0.1 to 0.3 lower than those of OEM).

Additionally, in the conditions of examinee-centred selection for the 2PL and 3PL models, OEM showed aberrant pattern in the RMSEs of the a-parameter values: for the pretest items with large true a-parameters (i.e. $a = 2.0$), the RMSEs of the a-parameters generated by OEM *increase* as the sample size increases. A possible explanation of this aberrant pattern is the lack of information for the a-parameters with the essentially 'match-b' process resulted from the examinee-centred selection (Chang and Ying, 2008), as explained earlier in this paper. In contrast, MEM is less sensitive to this problem and was able to successfully stabilise the estimation: the RMSEs not only decrease as the sample size increases but also were significantly smaller than those generated by OEM.

Taking all factors into consideration, for the 1PL model, OEM is preferred because of the shorter computation time and similar level of accuracy compared to MEM; for the 2PL and 3PL models, MEM is preferred because of the superior and more stable estimation results; but if computation time is a concern, OEM is also a viable choice as long as the pretest items are not selected by the examinee-centred method, and especially not all items have large true a-parameter values. Based on these findings, the subsequent analysis of the other factors was made using the results from the MEM estimation.

5.2 Comparing seeding locations

Effects of seeding locations were found in the 1PL model in items with more extreme b-parameter values (i.e. $b = -2.0, -1.5, 1.0, 1.5, 2.0$). For those items, as shown in Figure 3[1], the estimation accuracies of examinee-centred item selection and the OIRPI method are generally better in the middle and late seeding locations than in the early seeding location. This effect of seeding location diminishes for those moderate b-parameter values. A possible explanation for this phenomenon is when items with more extreme true b-parameter value are matched with examinees based on highly inaccurate θ estimates due to the early seeding, the true θ values of those examinees may be very far away from the informative (i.e. matched by true parameter values) location, so very little information will actually be collected; while for items with more moderate parameter values, even if the matched examinees have true θ values that are somewhat away from the b-parameter value of the items, they will not be as far away as in the previous case, so that still a decent amount of information can be collected for them even with early seeding.

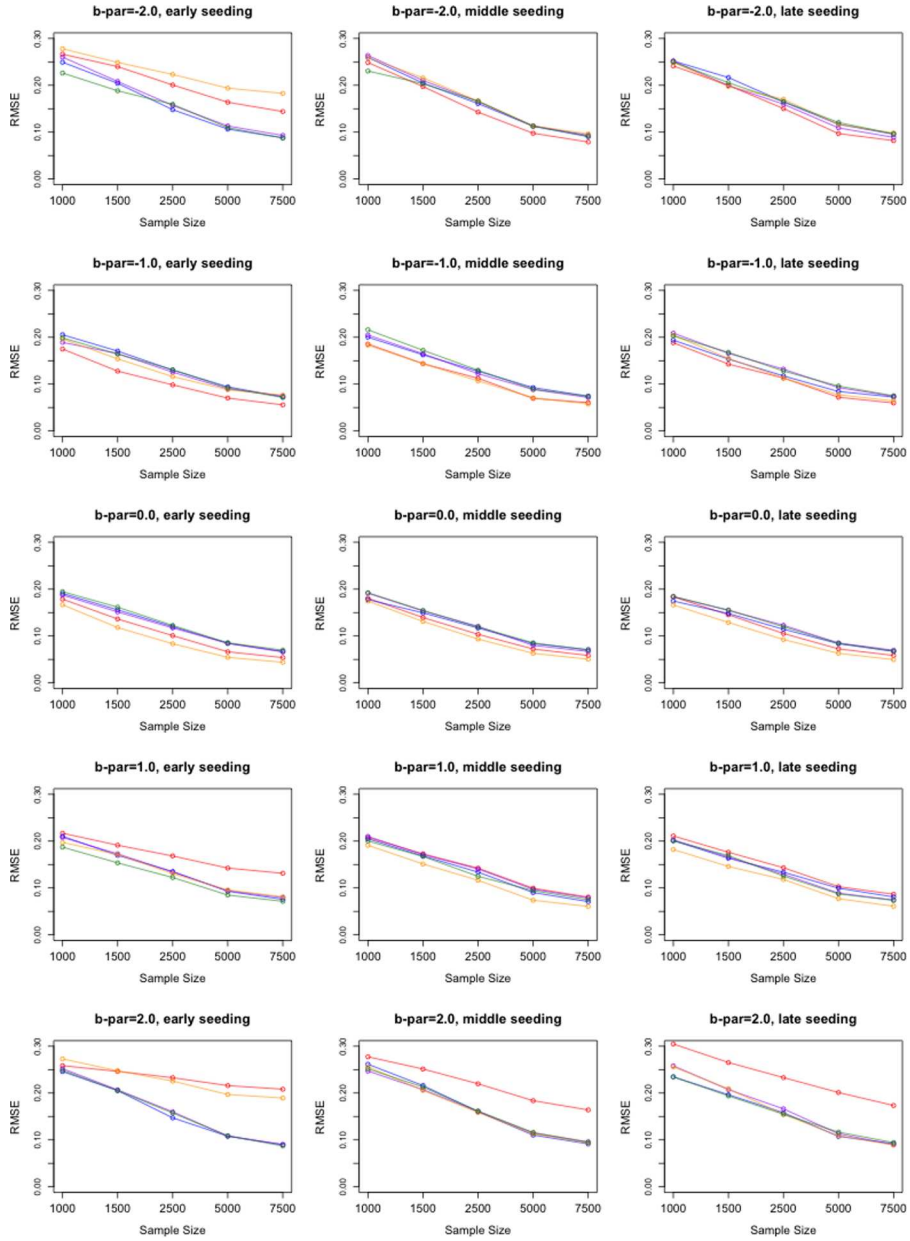
Effects of seeding locations were not quite visible for other item selection methods for the 1PL model and most conditions in the 2PL and 3PL models. This may suggest that overall seeding location does not make much difference in the eventual calibration outcome in the highly complicated online calibration system, which could provide test developers with more flexibility in randomly seeding pretest items throughout the entire test so that it is less predictable by the examinees, resulting in less contaminated calibration data.

5.3 Comparing calibration sample sizes

An obvious and consistent trend has been observed for different calibration sample size conditions for the 1PL model across most pretest item selection methods, seeding locations, and true a- and b-parameter values as shown in Figure 3. Calibration accuracy improves as the sample size increases. The rate of improvement is consistent from the recording point of 1000 examinees up to 5000 examinees and levels out from 5000 examinees to

7500 examinees. This means that the improvement in calibration accuracy is only marginal from an average of about 700 samples per pretest item to an average of about 1000 samples per pretest item in the 1PL model (the sample size conversion was explained in the Section 4.1).

Figure 3 The RMSE of *b*-parameter estimates under the 1PL model (see online version for colours)



Note: red=OIRPI, blue=B-D-optimal, purple=D-optimal, orange=Examinee, green=Random

In the 2PL and 3PL models, calibration accuracy for a- and b-parameters generally improves as the sample size increases but at varied rates, and the effect of sample size is greater on items with low a-parameters, as shown in Figures 4–7. In contrast, the calibration accuracy of c-parameters could decrease as the sample size increases, especially when a-parameters are low (i.e. at the 0.5 level), as shown in Figure 8. At other a-parameter values, calibration accuracy of c-parameters essentially stays the same. This is not surprising because calibration difficulty for the c-parameter is well known for the 3PL model (Swaminathan and Gifford, 1979).

5.4 Comparing pretest item selection methods

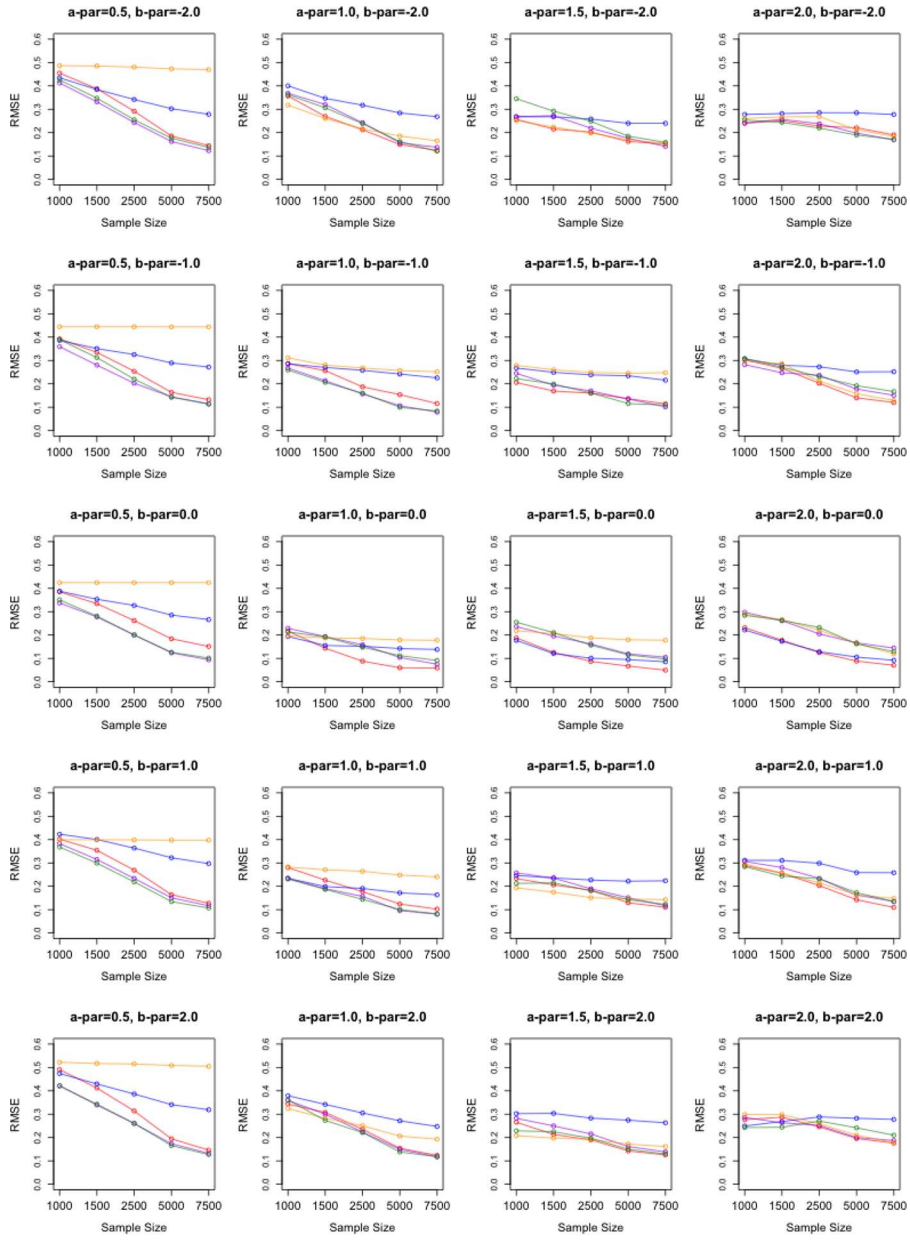
For the 1PL model, Figure 3 will be used to compare the performance of pretest item selection methods. For the 2PL and 3PL models, ICC area differences (as given by Eq. 18) were calculated to evaluate the overall accuracy of all two or three parameters and used as the single-number summary to compare the performance of pretest item selection methods. The results in ICC area differences are presented in Figures 9 and 10. Note that because the seeding locations do not have significant effect for the 2PL and 3PL models, for presentation simplicity, only the middle seeding location conditions are presented for the 2PL and 3PL models.

The pattern among pretest item selection methods depends on the true b-parameter values for the 1PL model and the true a-parameter values for the 2PL and 3PL models. As Figure 3 shows, under the 1PL model, for most conditions with moderate b-parameter values, the OIRPI method (red lines) and examinee-centred method (orange lines) generated slightly lower RMSE values than the other two, but the difference is very small. This trend reverses for the more extreme b-parameter values especially for the early seeding conditions, where random selection (green lines) and the Bayesian D-optimal method (blue lines) performed better.

The differences across pretest item selection methods are more obvious in the 2PL and 3PL models as shown in Figures 9 and 10. The studied adaptive item selection methods were only found slightly more efficient than the random selection method in the 2PL model for items with larger a-parameter values and smaller absolute b-parameter values. Alternatively, for items with $a = 0.5$, across most b-parameter values in the 2PL and 3PL models, the random selection method and the original D-optimal method generated the smallest ICC area difference, followed by the OIRPI method, then the Bayesian D-optimal method and lastly the examinee-centred method.

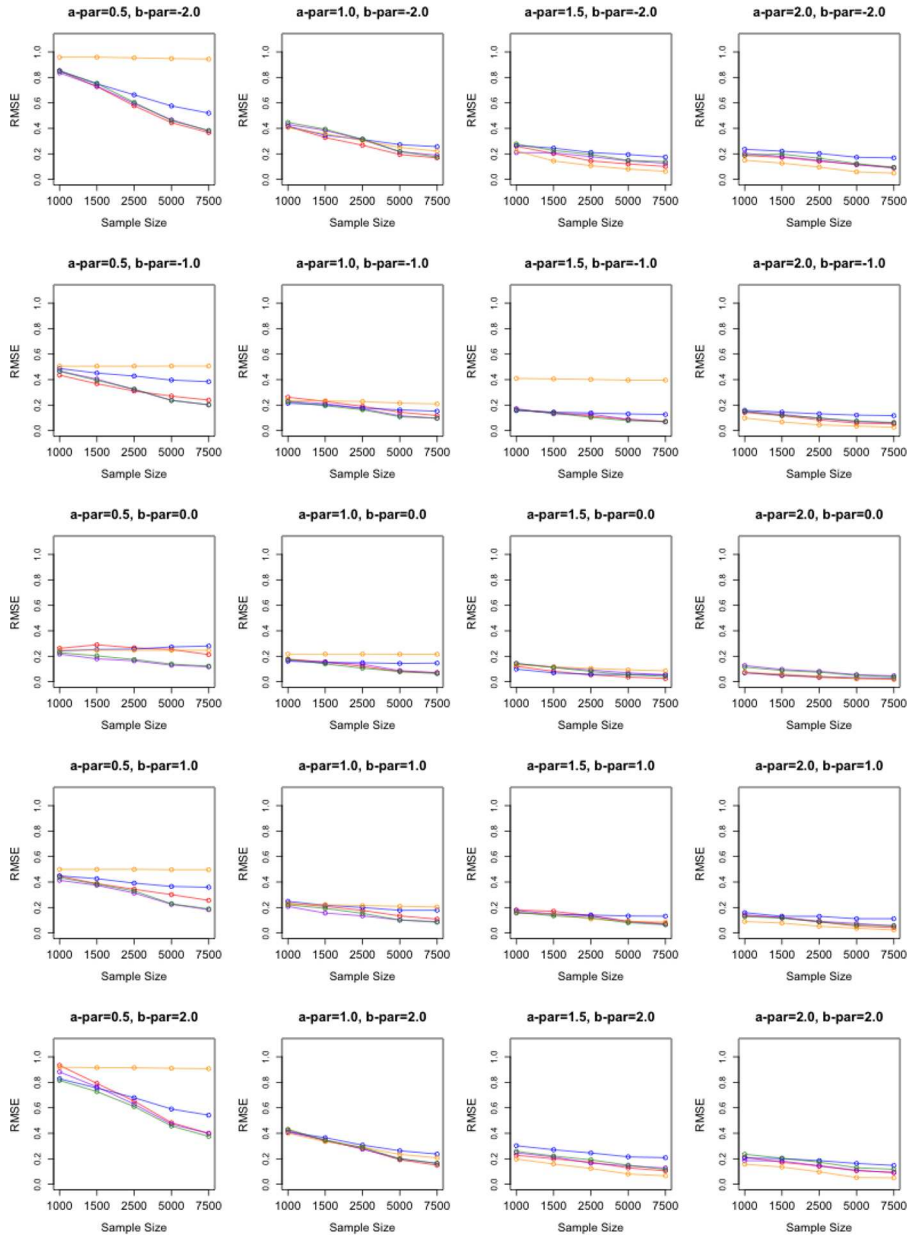
In summary, the calibration accuracy resulting from various pretest item selection methods seems to depend on the value of the a- and b-parameters. However, overall, the simplest random selection seems to be comparable or even more efficient than the other more complicated adaptive item selection methods. This result might be counterintuitive, but this finding could indicate that in the sophisticated setting of online calibration, random selection may have been demonstrated an ideal choice after all.

Figure 4 The RMSE of a-parameter estimates under the 2PL model with middle seeding location (see online version for colours)



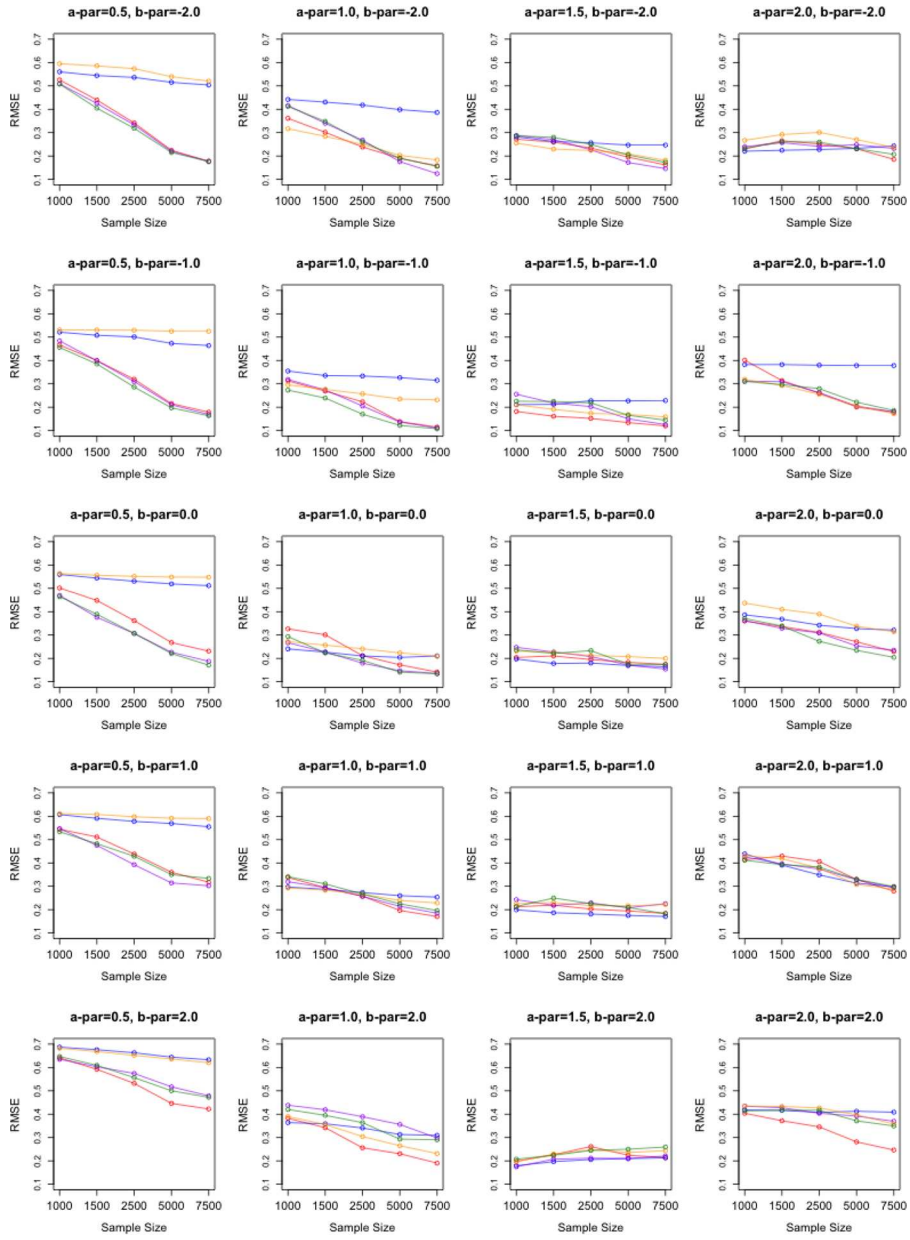
Note: red=OIRPI, blue=B-D-optimal, purple=D-optimal, orange=Examinee, green=Random

Figure 5 The RMSE of *b*-parameter estimates under the 2PL model with middle seeding location (see online version for colours)



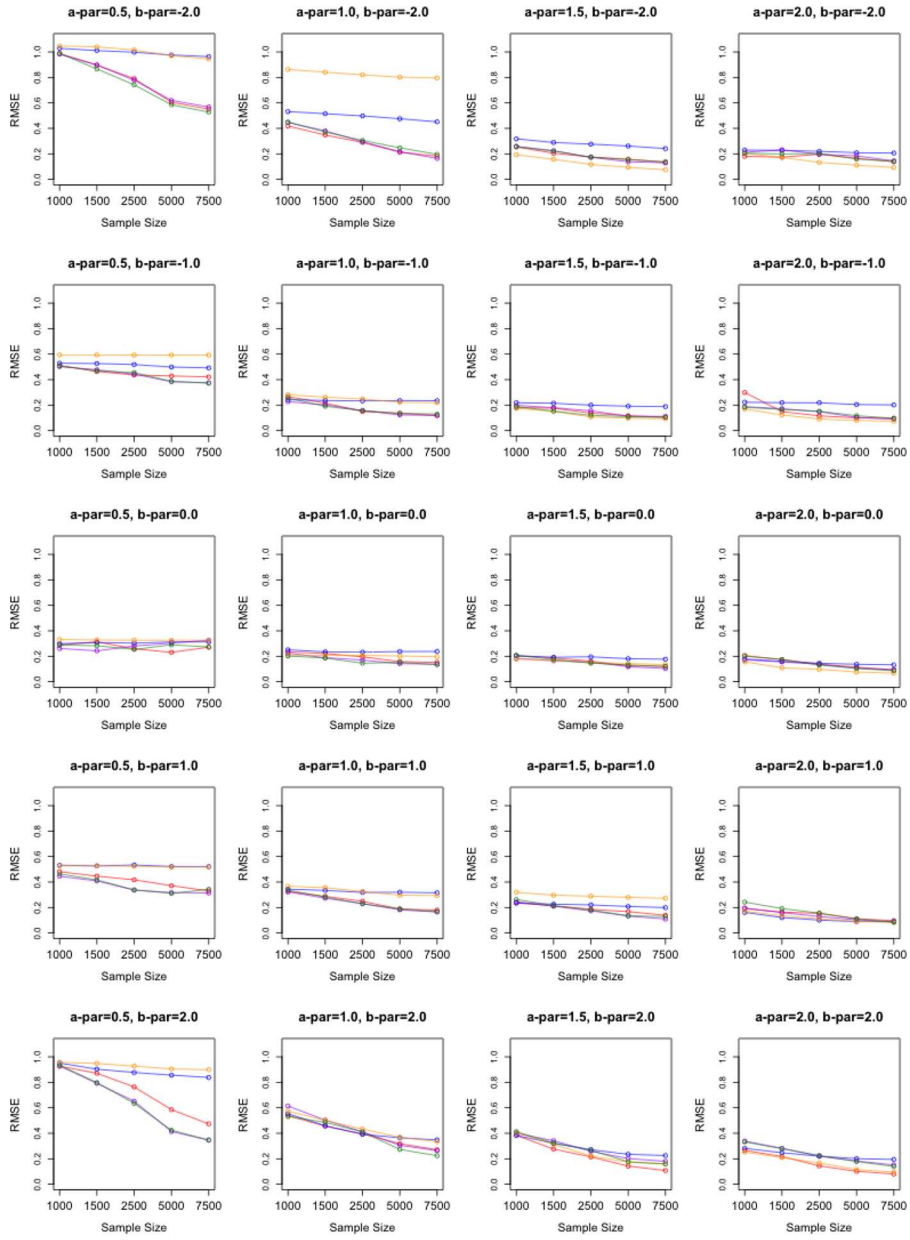
Note: red=OIRPI, blue=B-D-optimal, purple=D-optimal, orange=Examinee, green=Random

Figure 6 The RMSE of a-parameter estimates under the 3PL model with middle seeding location (see online version for colours)



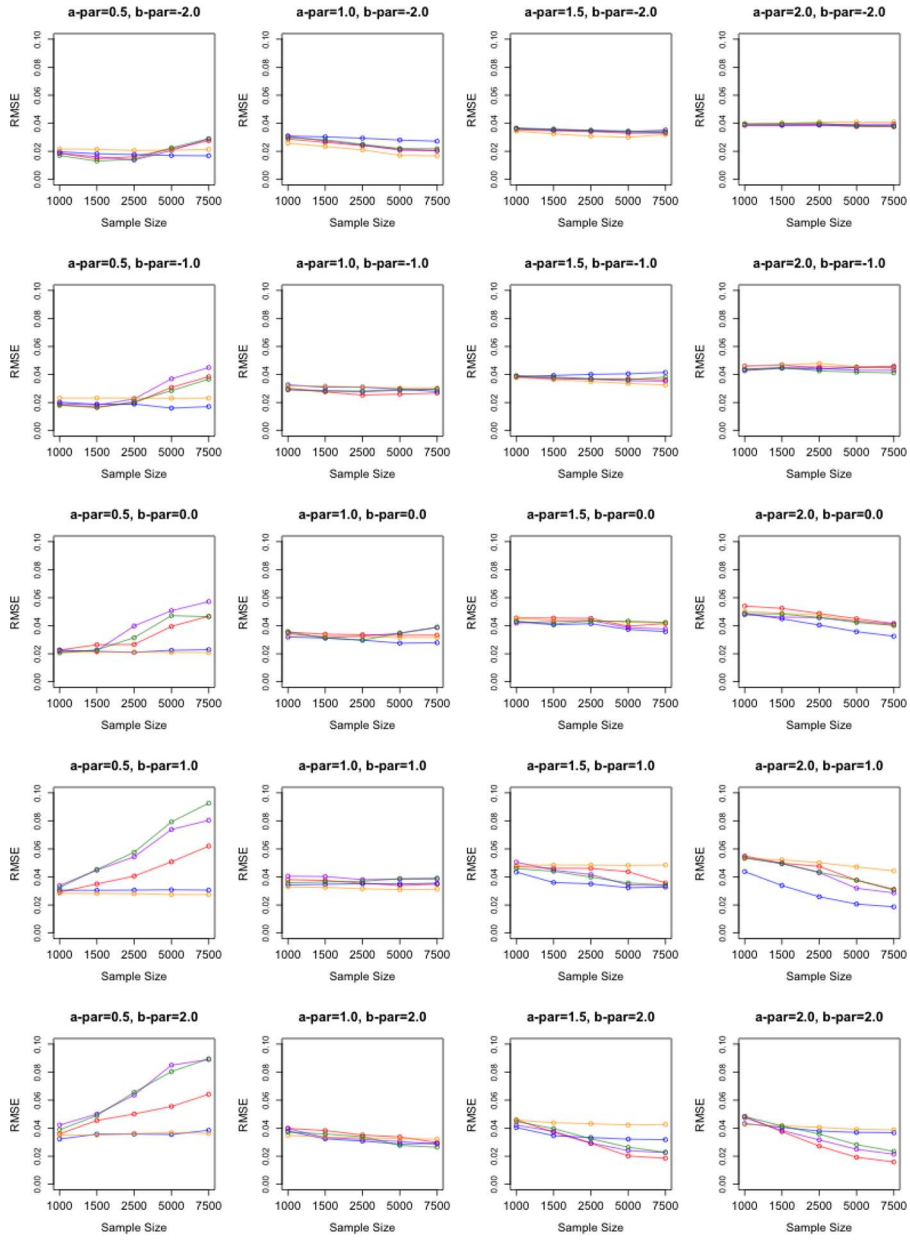
Note: red=OIRPI, blue=B-D-optimal, purple=D-optimal, orange=Examinee, green=Random

Figure 7 The RMSE of *b*-parameter estimates under the 3PL model with middle seeding location (see online version for colours)



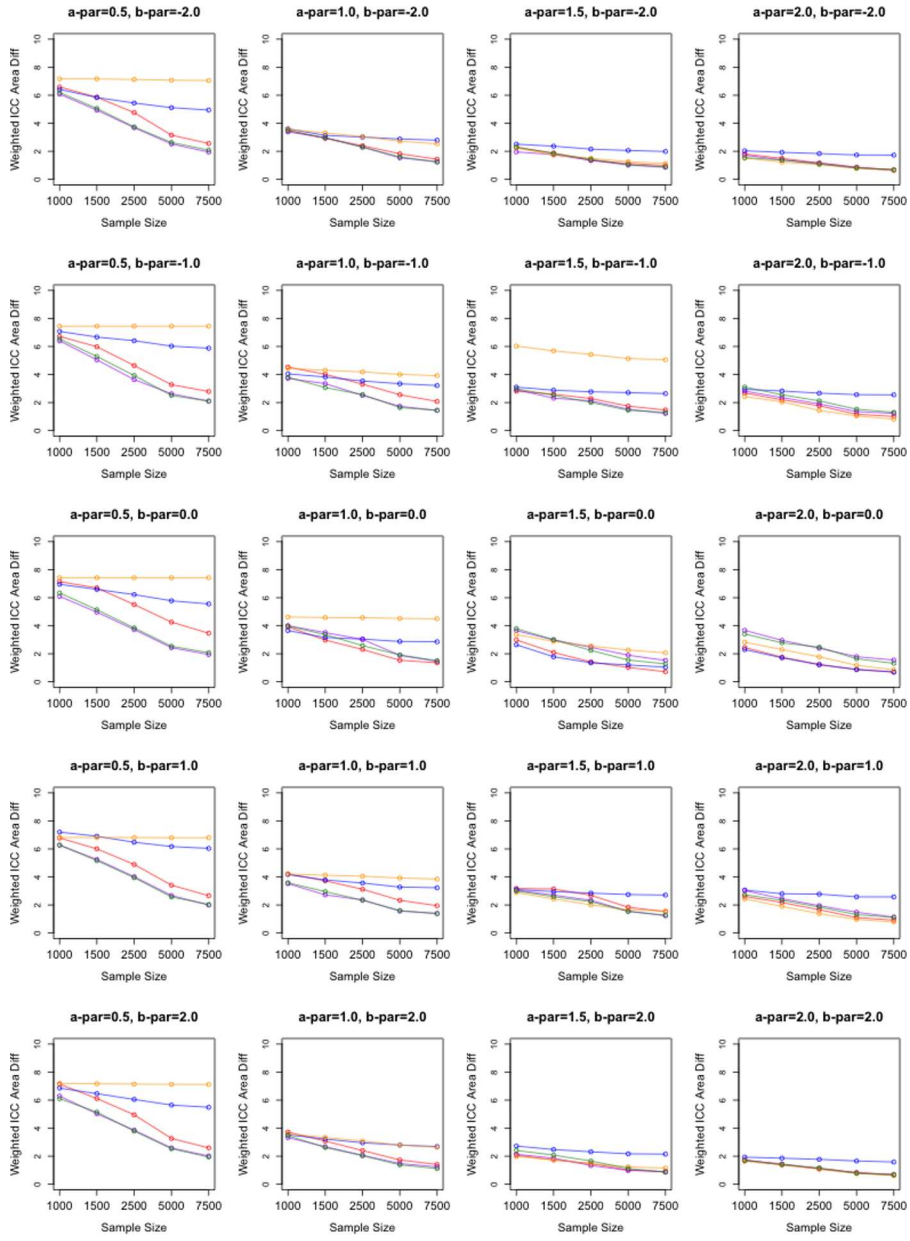
Note: red=OIRPI, blue=B-D-optimal, purple=D-optimal, orange=Examinee, green=Random

Figure 8 The RMSE of *c*-parameter estimates under the 3PL model with middle seeding location (see online version for colours)



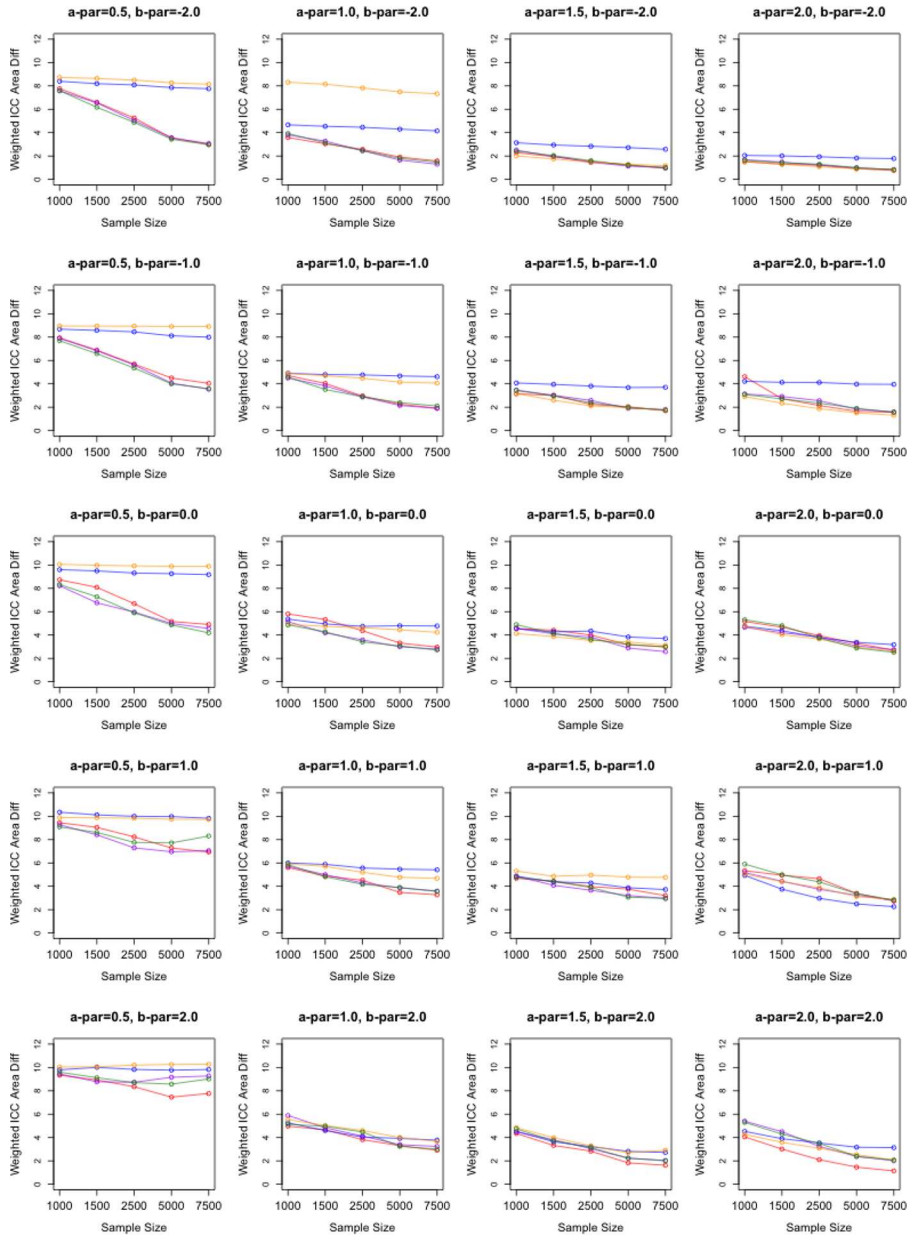
Note: red=OIRPI, blue=B-D-optimal, purple=D-optimal, orange=Examinee, green=Random

Figure 9 The ICC area difference under the 2PL model with middle seeding location (see online version for colours)



Note: red=OIRPI, blue=B-D-optimal, purple=D-optimal, orange=Examinee, green=Random

Figure 10 The ICC area difference under the 3PL model with middle seeding location (see online version for colours)



Note: red=OIRPI, blue=B-D-optimal, purple=D-optimal, orange=Examinee, green=Random

6 Discussion

A strong trend towards the use of technology-enhanced assessments has been seen in recent years. Many educational and psychological assessments are moving towards computerised (adaptive) modes, such as the new K-12 state assessments created under the Race To The Top (RTTT) program,[2] the National Assessment of Educational Progress (NAEP),[3] and the patient-reported outcome (PRO) measurements in medical practices (Fayers, 2007; Zheng et al., 2013). Many of these CAT programs are high stakes and administered over multiple years, and therefore, this places high demands on item calibration.

Online calibration has been studied for decades to dynamically sample examinees for calibrating new items more efficiently than the traditional pretesting methods. However, the existing methods for pretest item selection methods follow distinctive mechanisms and there has not been comparison among the options. This paper reported the findings from a simulation study that compared the performance of four available pretest item selection methods under the 1PL, 2PL and 3PL models and across varied conditions in terms of estimation methods, seeding locations and calibration sample sizes. It was found that the calibration accuracy resulting from various pretest item selection methods seems to depend on the value of the a- and b-parameters. And overall, the simplest random selection seems to be comparable or even more efficient than the other more complicated adaptive item selection methods. This finding could indicate that in the sophisticated setting of online calibration, random selection may have been demonstrated an ideal choice after all. Moreover, other findings from this study include:

1 MEM was more stable and accurate than OEM, especially for items with large a-parameter values.

2 no effect of the seeding location was observed.

There are several limitations in the current study. One limitation is that the results and conclusion from this study are limited within the specific simulation design. For example, the simulated c-parameters were fixed at 0.2 whereas in reality there may be items with significantly smaller or larger c-parameter values. Although an effort was made to better mimic a practical test setting, results under other settings still merit investigations, such as other test lengths, other operational item selection methods, or add content balancing and item exposure control, etc. Another limitation is that this simulation has not taken into consideration the possibility of differentiated performance if examinees identify the pretest items by the abrupt change in item difficulty compared to the overall adaptive trend. Real subject experiments may need to be conducted to answer this research question. Meanwhile, future studies may also be carried out to extend the current methods, such as adding termination rules or developing new methods for online calibration. With carefully developed online calibration designs, the item banks will hopefully be replenished or recalibrated more efficiently with more accurately calibrated items.

References

Anderson, T.W. (1984) *An Introduction to Multivariate Statistical Analysis*, 2nd ed., Wiley, New York, NY.

Armitage, P. (1975) *Sequential Medical Trials*, Blackwell, Oxford.

Baker, F.B. and Kim, S.H. (2004) *Item Response Theory: Parameter Estimation Techniques*, 2nd ed., Marcel Dekker, Inc., New York.

Ban, J.C., Hanson, B.A., Wang, T., Yi, Q. and Harris, D.J. (2001) 'A comparative study of on-line pretest item—calibration/scaling methods in computerised adaptive testing', *Journal of Educational Measurement*, Vol. 38, No. 3, pp.191–212.

Ban, J-C., Hanson, B.A., Yi, Q. and Harris, D.J. (2002) 'Data sparseness and on-line pretest item calibration-scaling methods in CAT', *Journal of Educational Measurement*, Vol. 39, No. 3, pp.207–218.

Berger, M.P.F. (1991) 'On the efficiency of IRT models when applied to different sampling designs', *Applied Psychological Measurement*, Vol. 15, No. 3, pp.293–306.

Berger, M.P.F. (1992) 'Sequential sampling designs for the two-parameter item response theory model', *Psychometrika*, Vol. 57, No. 4, pp.521–538.

Berger, M.P.F., King, J.C.Y. and Wong, W.K. (2000) 'Minimax D-optimal designs for item response theory models', *Psychometrika*, Vol. 65, No. 3, pp.377–390.

Buyske, S. (2005) 'Optimal design in educational testing', in Berger, M.P.F. and Wong, W.K. (Eds.): *Applied Optimal Designs*, John Wiley & Sons, Ltd., England, pp.1–19.

Chang, H-H. (2004) 'Understanding computerised adaptive testing', in Kaplan, D. (Ed.): *The Sage Handbook of Quantitative Methods for the Social Sciences*, Sage Publications, Inc., Thousand Oaks, CA, pp.117–133.

Chang, H-H., Qian, J. and Ying, Z. (2001) 'A-stratified multistage computerised adaptive testing with b blocking', *Applied Psychological Measurement*, Vol. 25, No. 4, pp.333–341.

Chang, H-H. and Ying, Z. (2008) 'To weight or not to weight? balancing influence of initial items in adaptive testing', *Psychometrika*, Vol. 73, No. 3, pp.441–450.

Chang, Y-C.I. (2011) 'Sequential estimation in generalised linear models when covariates are subject to errors', *Metrika*, Vol. 73, pp.93–120.

Chang, Y-C.I. and Lu, H-Y. (2010) 'Online calibration via variable length computerized adaptive testing', *Psychometrika*, Vol. 75, No. 1, pp.140–157.

Chen, P. and Wang, C. (2016) 'A new online calibration method for multidimensional computerized adaptive testing', *Psychometrika*, Vol. 81, No. 3, pp.674–701.

Chen, P., Xin, T., Wang, C. and Chang, H-H. (2012) 'Online calibration methods for the DINA model with independent attributes in CD-CAT', *Psychometrika*, Vol. 77, No. 2, pp.201–222.

Fayers, P.M. (2007) 'Applying item response theory and computer adaptive testing: the challenges for health outcomes assessment', *Quality of Life Research*, Vol. 16, No. S1, pp.187–194.

Hambleton, R.K., Swaminathan, H. and Rogers, H.J. (1991) *Fundamentals of Item Response Theory*, Sage Publications, Inc., Newbury Park, CA.

Jones, D.H. and Jin, Z. (1994) 'Optimal sequential designs for on-line item estimation', *Psychometrika*, Vol. 59, No. 1, pp.59–75.

Kim, S. (2006) 'A comparative study of IRT fixed parameter calibration methods', *Journal of Educational Measurement*, Vol. 43, No. 4, pp.355–381.

Kingsbury, G.G. (2009) 'Adaptive item calibration: A process for estimating item parameters within a computerized adaptive test', in Weiss, D.J. (Ed.): *Proceedings of the 2009 GMAC Conference on Computerized Adaptive Testing*, pp.1–15.

Linden, W.J. van der and Glas, C.A.W. (2000) 'Capitalization on item calibration error in adaptive testing', *Applied Measurement in Education*, Vol. 13, No. 1, pp.35–53.

Linden, W.J. van der and Ren, H. (2015) 'Optimal Bayesian adaptive design for test-item calibration', *Psychometrika*, Vol. 80, No. 2, pp.263–288.

Segall, D.O. (2003) 'Calibrating CAT pools and online pretest items using MCMC methods', The annual meeting of the National Council on Measurement in Education, 22–24 April 2003, Chicago, IL.

Silvey, S.D. (1980) *Optimal Design*, Chapman and Hall, London, New York.

Stocking, M.L. (1988) *Scale drift in online calibration* (Tech. Rep. No. RR-88-28-ONR). Educational Testing Service, Princeton, NJ.

Swaminathan, H., and Gifford, J.A. (1979) *Estimation of parameters in the three-parameter latent trait model*, (Tech. Rep. No. RR-90), Amherst, MA.

Wainer, H. (2000) *Computer Adaptive Testing: A Primer*, Lawrence Erlbaum, Hillsdale, NJ.

Wainer, H. and Mislevy, R.J. (2000) 'Item response theory, item calibration, and proficiency estimation', Wainer, H. (Ed.): *Computer Adaptive Testing: A Primer*, Lawrence Erlbaum, Hillsdale, NJ, pp.65–102.

Wingersky, M.S. and Lord, F.M. (1984) 'An investigation of methods for reducing sampling error in certain IRT procedures', *Applied Psychological Measurement*, Vol. 8, No. 3, pp.347–364.

Ying, Z. and Wu, C.F.J. (1997) 'An asymptotic theory of sequential designs based on maximum likelihood recursions', *Statistica Sinica*, Vol. 7, pp.75–92.

Zheng, Y. (2014) *New Methods of Online Calibration for Item Bank Replenishment*, Unpublished doctoral dissertation, University of Illinois at Urbana-Champaign.

Zheng, Y., Chang, C-H. (2013) 'Content-balancing strategy in bifactor computerized adaptive patient-reported outcome measurement', *Quality of Life Research*, Vol. 22, No. 3, pp.491–499.

Zhu, R. (2006) *Implementation of Optimal Design for Item Calibration in Computerized Adaptive Testing (CAT)*. Unpublished doctoral dissertation, University of Illinois at Urbana-Champaign.

Notes

1 To achieve better visibility, only conditions with $b = -2.0, -1.0, 0, 1.0, 2.0$ are included in the figures.

2 See http://www.k12center.org for up-to-date developments.

3 See http://nces.ed.gov/nationsreportcard/about/future_of_naep.aspx.