Online Calibration of Polytomous Items Under the Generalized Partial Credit Model

Applied Psychological Measurement
2016, Vol. 40(6) 434–450
© The Author(s) 2016
Reprints and permissions:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/0146621616650406
apm.sagepub.com

Yi Zheng¹

Abstract

Online calibration is a technology-enhanced architecture for item calibration in computerized adaptive tests (CATs). Many CATs are administered continuously over a long term and rely on large item banks. To ensure test validity, these item banks need to be frequently replenished with new items, and these new items need to be pretested before being used operationally. Online calibration dynamically embeds pretest items in operational tests and calibrates their parameters as response data are gradually obtained through the continuous test administration. This study extends existing formulas, procedures, and algorithms for dichotomous item response theory models to the generalized partial credit model, a popular model for items scored in more than two categories. A simulation study was conducted to investigate the developed algorithms and procedures under a variety of conditions, including two estimation algorithms, three pretest item selection methods, three seeding locations, two numbers of score categories, and three calibration sample sizes. Results demonstrated acceptable estimation accuracy of the two estimation algorithms in some of the simulated conditions. A variety of findings were also revealed for the interacted effects of included factors, and recommendations were made respectively.

Keywords

online calibration, computerized adaptive testing, polytomous IRT models, generalized partial credit model

Invented about three decades ago (Stocking, 1988), online calibration is a technology-enhanced item calibration framework that pretests and calibrates new items on the fly during the continuous administration of *computerized adaptive tests* (CATs; Chang, 2014). Traditionally, CAT item banks are constructed using items calibrated from a separate pretesting event, usually administered under the paper-and-pencil testing mode, and there is a concern over the differential functioning of the items from the paper-and-pencil pretest to computerized adaptive operational mode. Now with the online calibration framework, after a CAT begins to operate using the initial item bank, new items can then be embedded inconspicuously within the operational tests and be pretested and calibrated within the same testing environment. This not only is

¹Arizona State University, Tempe, USA

Corresponding Author:

Yi Zheng, Arizona State University, 1050 S. Forest Mall, Tempe, AZ 85281, USA.
Email: yi.isabel.zheng@asu.edu

cost-effective but also enhances the consistency in both the mode of testing and examinee motivation.

Another important feature of online calibration is that it offers great flexibility in individual treatment for each pretest item. Analogous to what individual examinees experience in an operational CAT, any single pretest item can enter the online calibration process at any time, be assigned to a specific sample of examinees, and exported from the calibration phase flexibly. During the operational test, the parameter values of the pretest items are constantly updated, based on which the sampling scheme can be dynamically adjusted. Online calibration is in fact an application of *optimal design* in the education setting (Berger & Wong, 2005).

Traditional statistical parameter estimation algorithms in *item response theory* (IRT) have been adapted into the novel setting of online calibration (e.g., Ban, Hanson, Wang, Yi, & Harris, 2001; Ban, Hanson, Yi, & Harris, 2002; Stocking, 1988; Wainer & Mislevy, 2000). The most popular estimation methods proposed for online calibration are the *one EM cycle method* (OEM; Wainer & Mislevy, 2000) and the *multiple EM cycle method* (MEM; Ban et al., 2001). These algorithms make use of the existing parameter values of the operational items and naturally put the calibrated parameter values of the new items on the existing IRT scale without linking procedures, which is a highly favored feature. However, the parameter estimation algorithms for online calibration have only been developed and studied for dichotomously scored items such as multiple-choice questions and true/false questions. Online calibration of polytomously scored items, that is, items scored in more than two categories, is yet to be explored.

Polytomously scored items are becoming increasingly important in educational assessments due to the growing emphasis on non-multiple-choice items such as performance-based items. In the past, multiple-choice items have been the most common item type in large-scale tests because of their ease in scoring and conducting other psychometric procedures. Nowadays, computers and other digital devices have made rich response data readily available, such as click streams and the corresponding time duration of examinees' interaction with computers, and will continue to generate more and more diverse and informative "big data." This allows for and also encourages many new item types to be incorporated in current and future assessments. Consequently, psychometric procedures must also be extended to these new item types, and one goal of this article is to extend current online calibration methods to polytomous IRT models. A number of polytomous IRT models have been developed in recent decades. This article will discuss how the OEM method and MEM method can be extended to one of the most commonly used polytomous IRT models, the *generalized partial credit model* (GPCM; Muraki, 1992).

This article first presents a detailed description of the procedures, algorithms, and formulae for online calibration both in general and specific for polytomous items. This offers extensive information to researchers who are interested in replicating or extending the study and practitioners who are interested using the methods. Then a simulation study is presented, which examined the calibration results from the two calibration methods under a set of conditions fully crossed by three pretest item selection methods (random selection [RAND], maximum θ information selection [MTFI], and match- b selection [MATB]), three seeding locations (early, middle, and late in the test), and three sample sizes, which will be explained in detail shortly.

General Workflow of Online Calibration

The general procedures of online calibration can be summarized by Figure 1. First, a pretest item bank is formed by items to be calibrated. Then, during the operational test, when an examinee reaches a *seeding location* (i.e., an item position in the operational test allocated for pretest items), a pretest item is selected from the pretest item bank based on certain item selection

Figure 1. Flowchart of online calibration.

Note. CATs = computerized adaptive tests.

rules. Examinees' responses to these pretest items are not included in the scoring procedure, but only used for calibrating the item parameters. When the examinee finishes the test, the parameters of those administered pretest items are updated through statistical estimation. Sampling and parameter update are repeated for every new examinee or every new batch of examinees, where the sampling parameters are constantly adjusted based on the updated item parameter values. The sampling process of each pretest item can be terminated and the item can be exported from the pretest item bank individually once a satisfactory accuracy of the parameter estimation is achieved or the maximum sample size is reached. Finally, the exported pretest items are reviewed and, if approved, put into operational use.

Computation Logistics of Online Calibration

Because online calibration embeds item calibration in the continuous administration of the operational CAT, its data handling logistics and algorithms are different from those used for traditional item calibration with "static" response data. In traditional item calibration, all necessary response data are collected in full as the first step, and then a response matrix, which can be sparse sometimes, is prepared. This response matrix is then entered into a calibration program, which outputs the calibrated parameter values for all items of interest. In contrast, online calibration is an evolving process. More response data accumulate as more examinees take the test sequentially. To facilitate a sequential optimal design, the parameters of a pretest item need to be updated each time a new data point is received (termed *fully sequential*) or each time a batch of new data points is received (termed *group sequential*). Another complexity of online calibration is that each pretest item is taken by a different sample of examinees, and these examinees also take different sets of operational items. To calibrate any pretest item, data from all examinees who received this item are needed, which include both their responses to this item and their responses to all the varied operational items they took.

To facilitate online calibration, as each incoming examinee finishes the test, the following data must be stored in the database immediately. First, for each pretest item j that the examinee

received, assuming item j has already accumulated $N_j - 1$ samples, the N_j th entry of the following variables need to be stored for item j :

- i : The ID of the examinee;
- x_{ij} : The examinee's response to this pretest item j ;
- $t_i = (t_{i1}, t_{i2}, \dots, t_{iT_i})$: The IDs of the T_i operational items examinee i received; and
- $y_i = (y_{i1}, y_{i2}, \dots, y_{iT_i})$: The examinee's responses to these operational items.

The item parameters of all the operational items in the operational item bank, denoted by β_{op} , are also stored in the database and will be pulled when a pretest item is being calibrated.

If a fully sequential design is used, every time an examinee finishes the test, all pretest items administered to him or her will be estimated one by one. For a pretest item j being estimated, all necessary data from the previous $N_j - 1$ samples will be pulled from the database and used in combination with the new data from the current examinee to estimate the item parameter values. If a group sequential design is used, every time a pretest item obtains a certain amount of new response data, its parameter values are updated. The updated item parameter values will be used in the adaptive selection of pretest items for incoming examinees.

The GPCM

A variety of polytomous IRT models have been developed to model item response data scored in more than two categories. Among the available models, the GPCM (Muraki, 1992) is especially appropriate for and, thus, popular in education settings. More details on polytomous IRT models are given in online appendices.

As described by Muraki (1992), GPCM was developed based on the assumption that the conditional probability of selecting the k th category over the $(k - 1)$ th category, denoted by $P_{j,k|k-1}(\theta)$, is governed by the dichotomous response model:

$$P_{j,k|k-1}(\theta) = \frac{P_{jk}(\theta)}{P_{jk}(\theta) + P_{j,k-1}(\theta)} = \frac{\exp[a_j(\theta - b_{jk}^*)]}{1 + \exp[a_j(\theta - b_{jk}^*)]}. \quad (1)$$

Assuming there are a total of m_j score categories in item j , with mathematical derivation, the probability of a score in a particular category k , also termed as the *item category response functions* (ICRFs), can be directly expressed by

$$P_{jk}(\theta) = \frac{\exp\left[\sum_{v=1}^k a_j(\theta - b_{jv}^*)\right]}{\sum_{c=1}^{m_j} \exp\left[\sum_{v=1}^c a_j(\theta - b_{jv}^*)\right]}, \quad (2)$$

where b_{j1}^* is by convention set to 0 but can be any value without affecting the model outcomes. It is worth noting that the b_{jk}^* 's are the points on the θ scale where the ICRFs of the k th and $(k - 1)$ th categories intersect, and that b_{jk}^* 's are not necessarily sequentially ordered within item j although the response categories themselves need to be ordered, because the b_{jk}^* 's represent the relative probabilities between adjacent categories.

After fusing with Rating Scale Model (RSM) (Andrich, 1978), Muraki (1992) finally formulated GPCM into the following ICRF, which was used in this simulation study:

$$P_{jk}(\theta) = \frac{\exp\left[\sum_{v=1}^k a_j(\theta - b_j + d_{jv})\right]}{\sum_{c=1}^{m_j} \left[\sum_{v=1}^c a_j(\theta - b_j + d_{jv})\right]}, \quad (3)$$

where $d_{j1} \equiv 0$ for the same reason as above.

OEM and MEM: Marginal Maximum Likelihood Estimation (MMLE) With Expectation-Maximization (EM) Algorithm

With necessary examinee response data and operational item parameter data, a variety of estimation algorithms can be utilized to generate item parameter estimates for pretest items. This article focuses on two popular algorithms: OEM and MEM. Both OEM and MEM are based on the framework of MMLE with the EM algorithm. This framework integrates out (i.e., marginalizes) the examinee ability parameter θ using the posterior distribution of θ obtained from the response data and, then, it finds the item parameter estimates that maximize this marginalized log likelihood (Baker & Kim, 2004; Zheng, 2014). Due to page limit, the detailed formulae for the two estimation algorithms are described in the online appendices.

Simulation Study Design

A simulation study was conducted using a Fortran program written by the author to investigate the performance of the online calibration procedures and algorithms under various conditions. The program simulated the entire online calibration workflow including the operational CAT and the sequential sampling and calibration of the pretest items, replicated for 100 times. The study included five factors: estimation method (OEM and MEM), number of score categories (3 and 4), pretest item selection method (RAND, MTFI, and MATB), seeding location (early, middle, late in the test), and calibration sample size (approximately 200, 500, 1,000).

Generation of Item Parameters

In each replication, 300 operational items were randomly generated under the GPCM model from the following distributions:

$$a_j \sim \text{lognormal}(0.4, 0.1), \quad (4)$$

$$b_{jk}^* \sim \text{normal}(0, 1). \quad (5)$$

The pretest items were fixed at 20 prechosen parameter combinations to examine conditional results for items with distinctive characteristics. The a -parameter had four levels: 0.5, 1.0, 1.5, and 2.0. They were crossed with five different combinations of the b_{jk}^* 's, which are listed and illustrated with category characteristic curves correspondingly in Figure 2.

As shown in Figure 2, in both the three-category and four-category conditions, the first combination of b_{jk}^* 's represents an item of moderate difficulty with balanced categories. The second/third combination represents an easy/hard item with all categories being easy/hard. The fourth combination represents an item whose beginning categories are easy but ending category is relatively hard. The fifth combination represents an item whose beginning category is relatively easy but ending categories are hard.

Figure 2. Category characteristic curves for the simulated pretest items with $a = 1.5$ and b_{jk}^* 's set at five different combinations.
Note. The left column items are for the three-category conditions, and the right column items are for the four-category conditions.

The b_{jk}^* 's were then converted into one location parameter b_j plus m_j threshold parameters, d_{jk} , as expressed by Equation 3. As mentioned earlier, d_{j1} can be of any arbitrary value without affecting the model outcome and therefore was set to 0 for convenience. In addition, for model identification purposes, d_{j2} was set to 0 as well. Only d_{j3} through d_{jm_j} were variable.

CAT Workflow

In each replication, 4,000 examinee ability θ 's were randomly generated from the standard normal distribution. In the simulation, examinees took the test sequentially. The test length was 40 and among the 40 items, 35 were operational items and five were pretest items.

During the test, operational items were selected from the operational item bank using the maximum Fisher information method, a typical operational item selection method in CAT. The Fisher information of a GPCM item j is formulated in Muraki (1993) as below:

$$I_j(\theta) = \sum_{k=1}^{m_j} P_{jk}(\theta) \left[-\frac{\partial^2}{\partial \theta^2} \ln P_{jk}(\theta) \right], \quad (6)$$

$$= a_j^2 \sum_{k=1}^{m_j} (T_k - \bar{T}_j(\theta))^2 P_{jk}(\theta), \quad (7)$$

where $T_k = 1, 2, 3, \dots$ are the category scores and $\bar{T}_j(\theta) = \sum_{k=1}^{m_j} T_k P_{jk}(\theta)$. During operational item selection, provisional θ estimates were used to replace the θ 's in the formulae. After each operational item was administered, the examinee ability parameter $\hat{\theta}$ was updated using the *expected a-posteriori* (EAP) method (Baker & Kim, 2004).

When an examinee reaches a seeding location, one of the pretest items in the pretest item bank was selected and administered. Examinees' responses to pretest items were not used to update $\hat{\theta}$. The item parameters of each pretest item were updated after it obtained every 10 new samples. Because parameter estimation is highly unstable when the sample size is too small, a random item selection phase was included in the MTFI and MATB conditions to initialize the provisional pretest item parameters. Specifically, for the first 400 examinees, which means on average the first $400 \times 5/20 = 100$ responses each pretest item receives, pretest items were selected randomly and the item parameters were not updated. After that, the adaptive selection was used. A pretest item was exported from the pretest item bank when it received more than 1,000 samples.

Pretest Item Selection Methods

Three pretest item selection methods were included in this simulation study: RAND, MTFI, and MATB. In RAND, one pretest item was randomly selected from the pretest item pool when an examinee reaches a seeding location.

In MTFI, the pretest item that maximizes the Fisher information for $\hat{\theta}$ was selected. This is the same method as used in selecting operational items in this simulation study as explained in the previous section, with the only difference being that item parameters in Equations 6 and 7 were known for operational items, whereas for pretest items, provisional item parameter estimates were used. Zheng (2014) termed selecting pretest items using the same method as in selecting operational items "examinee-centered item selection," and this examinee-centered selection approach has been used in previous studies of online calibration (e.g., Chen, Xin, Wang, & Chang, 2012; Kingsbury, 2009). An advantage of this approach is that the pretest items will blend in the surrounding operational items so the consistency in testing environment and examinee motivation can be well ensured.

MATB is a model-based heuristic item selection method. This method matches the mean of provisional estimates of b^* 's in Equation 2 with the provisional θ estimate of the incumbent examinee. Every time an examinee reaches a seeding location, the distance between his or her

current θ estimate and the mean of provisional estimates of b^* 's was computed for each pretest item, and the item with the shortest absolute distance was selected.

Seeding Locations

Although in the proposed online calibration workflow, pretest item parameters are not updated until the end of any examinee's test, seeding location may still influence calibration results when pretest items are selected adaptively, because adaptive item selection methods involve θ estimates, and θ estimates are of different levels of accuracy at different stages of the test.

Thus, seeding location was included as a factor in the simulation study, and three levels were chosen for seeding location: (a) early in the test (Items 6 through 10), (b) in the middle of the test (Items 19 through 23), and (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.

Calibration Sample Sizes

In this simulation study, item parameter estimates were recorded at three different time points when 800, 2,000, 4,000 examinees have taken the test. With random pretest item selection, given that each examinee takes five pretest items and there were a total of 20 pretest items, these chosen conditions corresponds to an average of roughly 200, 500, and 1,000 calibration samples per pretest item.

Evaluation Criteria

The simulation results were evaluated through the root mean square errors (RMSEs) of the estimates of each item parameter, formulated by Equation 8.

$$\text{RMSE}_p = \sqrt{\frac{1}{J} \sum_{j=1}^J (\hat{\beta}_p - \beta_p)^2}, \quad (8)$$

where p denotes the specific element in the item parameter vector, such as the a -parameter, b -parameter, and d_k -parameters. A smaller value means the estimation is more accurate.

Simulation Results

Results of the simulation study were averaged across the 100 replications and are presented in Figures 3 to 8.

Effect of Seeding Locations

Effects of seeding locations were almost unnoticeable from results of the simulation. In other words, graphs of the RMSEs of item parameter estimates look fairly similar for early, middle, or late seeding locations under various conditions with regard to other factors. This may suggest

Figure 3. The RMSE of the a -, b -, and d -parameters by random item selection with the late seeding location.

Note. Only those variable entries in the d_k parameters are included, and when there are more than one variable d_k parameters, the graph reflects the averages of the individual RMSE values for the d_k 's. RMSE = root mean square error; OEM = one EM cycle method; MEM = multiple EM cycle method.

Figure 4. The RMSE of the a -, b -, and d -parameters by maximum θ information selection with the late seeding location.

Note. Only those variable entries in the d_k parameters are included, and when there are more than one variable d_k parameters, the graph reflects the averages of the individual RMSEs values for the d_k 's. RMSE = root mean square error; OEM = one EM cycle method; MEM = multiple EM cycle method.

that, on 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

Figure 5. The RMSE of the a -, b -, and d -parameters by match- b selection with the late seeding location. Note. Only those variable entries in the d_k parameters are included, and when there are more than one variable d_k parameters, the graph reflects the averages of the individual RMSEs values for the d_k 's. RMSE = root mean square error; OEM = one EM cycle method; MEM = multiple EM cycle method.

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.

Because of the similarity among seeding locations, to provide a clearer and more focused presentation, only results from late seeding location are presented in the article. Results from early and middle seeding locations are given in online appendices.

Interacted Effects of Sample Sizes, Number of Categories, Estimation Algorithms, and Pretest Item Selection Methods

Figures 3 to 5 present the effect of calibration sample size under various conditions. As described in the "Simulation Study Design" section, item parameter estimates were recorded 3 times in each replication—when 800, 2,000, and 4,000 examinees finished the test. With random pretest item selection, these correspond to calibration sample sizes of about 200, 500, and 1,000 for each pretest item. For the other two item selection methods, the actual sample sizes can vary for different items. For all conditions, calibration accuracy (as reflected by low values in RMSE) increases as sample size increases, which empirically verifies the consistency of the estimation algorithms.

Across all conditions, 200 seems too small a sample size for a reasonable calibration accuracy. When the sample size gets to about 1,000 per item, all conditions seem to have reasonable and similar average RMSE values across all items. (Yet conditional results are different for individual items. See the next section for details.) When the sample size is about 500 per item, pretest item selection methods and estimation algorithms make a difference. For RAND with both OEM and MEM, and MATB with MEM, there was only small improvement in calibration accuracy from the sample size of 500 to 1,000, so 500 could be an economic sample size with reasonable calibration accuracy for these settings. However, for MATB with OEM, and MTFI with both OEM and MEM, RMSE with the sample size of 500 was still noticeably larger than with

sample size of 1,000. Therefore, it is recommended that appropriate calibration sample sizes should be chosen with respect to both the pretest item selection method and the parameter estimation approach.

The estimation of the three-category model was noticeably more accurate than the estimation of the four-category model for OEM conditions. For MEM conditions, the two models did not show noticeable difference. In both models, the estimation of a -parameters was more stable than the b - and d -parameters. There was difficulty in generating accurate b - and d -parameter estimates in the smaller sample size conditions, and this inaccuracy was more severe in OEM conditions than in MEM conditions.

Conditional RMSE Results for Individual Items

Figures 6 through 8 summarize RMSE values of each individual pretest item under various conditions when the average sample size was 1,000 per item. As explained earlier and shown in Figure 2, the 20 pretest items were generated by crossing four typical and distinctive a -parameter values and five typical and distinctive b_{jk}^* combinations. Due to page limit, only results for the three-category model conditions are presented in the article; results for the four-category model conditions, which are largely similar to the formers, are given in the online appendices.

Comparing calibration algorithms. Comparing OEM and MEM, when the sample size was as large as about 1,000 per item on average, OEM and MEM generated quite comparable RMSE values. Most of time, MEM was slightly more accurate than OEM. Considering that MEM takes considerably longer time than OEM in computation, OEM may be a viable option when sample size is large. But remember from Figures 3 to 5 that when sample size is small, MEM was significantly more accurate than OEM.

Comparing different items. Regarding a -parameters, a salient trend seen in Figure 6 is that calibration accuracy of a -parameters decreases as the true a -parameter increases. Items with a lower true a -parameter value have more accurate a -parameter estimates, and this is true across all conditions in other factors.

Regarding b - and d -parameters, there is an overall trend in Figures 7 and 8 that calibration accuracy of both b - and d -parameters often increases as the true a -parameter increases. However, this trend is fairly subtle, and there are occasional exceptions in various conditions.

Interesting observations emerge from the patterns in calibration accuracy of b - and d -parameters by different b_{jk}^* combinations. In both three- and four-category models, the first, third, and fifth b_{jk}^* combinations had more accurate b -parameter estimates—these correspond to items of moderate difficulty with balanced categories, hard items with all categories being hard, and items whose beginning category is relatively easy but ending categories are hard. Compared with them, the two groups of relatively easy items did not have as accurate b -parameter estimates. The pattern in d -parameter estimates is not as obvious, but noticeably, d -parameter estimates were more accurate for items of moderate difficulty with balanced categories and relatively high true a -parameter values.

Comparing pretest item selection methods. Comparing the three pretest item selection methods, the most common trend across the three investigated item selection methods is that RAND generated slightly more accurate item parameter estimates than MTFI, which is in turn slightly more accurate than MATB. This trend is reversed only in b - and d -estimates for the easiest items.

This seemingly counter-intuitive result can be further understood by examining the distributions of θ 's each item received, which are given in the online appendices. Because the simulated

Figure 6. The RMSE of the a -parameters at sample size of 1,000 per item for the three-category model with the late seeding location.

Note. RMSE = root mean square error; OEM = one EM cycle method; MEM = multiple EM cycle method; MATB = match- b selection; MTFI = maximum θ information selection; RAND = random selection.

Figure 7. The RMSE of the b -parameters at sample size of 1,000 per item for the three-category model with the late seeding location.

Note. RMSE = root mean square error; OEM = one EM cycle method; MEM = multiple EM cycle method; MATB = match- b selection; MTFI = maximum θ information selection; RAND = random selection.

Figure 8. The RMSE of the d -parameters at sample size of 1,000 per item for the three-category model with the late seeding location.

Note. Only those variable entries in the d_k parameters are included, and when there are more than one variable d_k parameters, the graph reflects the averages of the individual RMSEs values for the d_k 's. RMSE = root mean square error; OEM = one EM cycle method; MEM = multiple EM cycle method; MATB = match- b selection; MTFI = maximum θ information selection; RAND = random selection.

population θ distribution is normal, RAND leads to an approximately normal distribution of θ for each item. For MTFI and MATB, the distributions of θ received by each item were skewed, respectively.

The inferior performance of MTFI is no surprise because the operational item selection method, which is the maximum Fisher information method in this study, is typically designed to optimize the measurement accuracy for the examinee ability parameter θ but not designed to optimize the calibration accuracy of the item parameters (Zheng, 2014). With MTFI, there was even one case (Figure 7 row 4 column 4) where the estimation did not seem to converge.

These results also suggest that a more spread θ distribution may generate more accurate estimates for all item parameters than a skewed θ distribution, consistent with Stocking's (1990) findings for the three-parameter logistic model. This claim is further supported by a closer examination of the θ distribution figures in the online appendices in conjunction with Figures 6 through 8. MATB generally led to more skewed θ distributions than MTFI for items in the first, fourth, and fifth rows, and for these same items RMSE was larger in MATB than in MTFI.

Discussion

This study extended two estimation algorithms for online calibration, OEM and MEM, which were originally developed for dichotomous IRT models, to the GPCM, a popular IRT model for polytomously scored items. Detailed description of the algorithms and the unique computation logistics of online calibration were given in the article, which may assist researchers and practitioners in applying these methods or replicating or extending the study.

A simulation study was conducted to investigate the performance of the developed OEM and MEM algorithms under a variety of conditions. Results showed that both OEM and MEM were able to generate reasonably accurate parameter estimates with a sample size of about 1,000 per item. MEM was more accurate than OEM when sample size was smaller, but took longer to complete because of the multiple EM cycles.

This study also revealed the following key findings. (a) All seeding locations generated similar results. Thus test developers may have more flexibility to embed pretest items throughout the entire test. (b) When OEM was used, parameter estimation was noticeably more accurate for the three-category model than the four-category model; when MEM was used, the two models do not show noticeable difference. (c) In most conditions RAND generated more accurate parameter estimates than MTFI and MATB, and this may be attributed to the difference in distribution of θ received by each item—a more spread θ distribution led to more accurate item parameter estimation than a skewed θ distribution for GPCM. (d) Recommendations for minimum calibration sample sizes were made for specific conditions with regard to other design factors.

Cautions should be used in interpreting or generalizing the findings from the study, because the findings are limited by the specific test design, calibration design, numbers and distributions of the simulated item parameters, θ estimation algorithm, operational item selection method, and so on.

Also note that the item parameters calibrated through online calibration not only can be used in subsequent operational CAT tests, but they may also be useful in assembling tests in other formats, including linear tests (e.g., Veldkamp, Matteucci, & de Jong, 2013) and multistage tests (e.g., Zheng, Wang, Culbertson, & Chang, 2014). If format difference is of concern in using these item parameters for *scoring*, the items can be re-calibrated with response data collected in the actual delivery format, but still these item parameters obtained from online calibration can provide reasonable approximation in the test *assembly* step.

There is a great need for more research in online calibration. For example, more research is needed to investigate the methods in similar or new settings to corroborate or compare with the findings from this article. In addition, this article revealed many interesting empirical patterns in conditional results—more theoretical work is needed to find explanations for these patterns.

Furthermore, because neither of the two adaptive pretest item selection methods investigated in this study showed superior performance than RAND, it is of great interest to develop optimal-design-based, “item-centered” pretest item selection methods (Zheng, 2014; Zheng & Chang, 2014) for polytomous IRT models to hopefully better capitalize on the power of online calibration. Here item-centered pretest item selection methods refer to those that rely on criteria derived for item parameter estimation, such as the D-optimal criterion (i.e., determinant of item parameter information matrix), as opposed to examinee-centered criteria, which are derived for examinee parameter estimation, such as maximum θ information method (van der Linden & Ren, 2015; Zheng, 2014).

The algorithms discussed in this article can also be extended to other polytomous IRT models. Also along this line are termination rules—termination rules originally developed for multi-dimensional CAT (Wang, Chang, & Boughton, 2013) may be adapted into online calibration settings. Finally, using online calibration to build vertical scales (Li & Lissitz, 2012; Tong & Kolen, 2007) and detect item parameter drift (Babcock & Albano, 2012; Guo, Zheng, & Chang, 2015) remains promising future directions.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

Supplemental Material

The online appendices are available at <http://apm.sagepub.com/supplemental>

References

- Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, 43, 561-573.
- Babcock, B., & Albano, A. D. (2012). Rasch scale stability in the presence of item parameter and trait drift. *Applied Psychological Measurement*, 36, 565-580. doi:10.1177/0146621612455090
- Baker, F. B., & Kim, S.-H. (2004). *Item response theory: Parameter estimation techniques* (2nd ed.). New York, NY: Marcel Dekker.
- Ban, J. C., Hanson, B. A., Wang, T., Yi, Q., & Harris, D. J. (2001). A comparative study of on-line pretest item—Calibration/scaling methods in computerized adaptive testing. *Journal of Educational Measurement*, 38, 191-212. doi:10.1111/j.1745-3984.2001.tb01123.x
- Ban, J.-C., Hanson, B. A., Yi, Q., & Harris, D. J. (2002). Data sparseness and on-line pretest item calibration-scaling methods in CAT. *Journal of Educational Measurement*, 39, 207-218. doi: 10.1111/j.1745-3984.2002.tb01174.x
- Berger, M. P., & Wong, W. K. (2005). *Applied optimal designs*. Chichester, UK: John Wiley.
- Chang, H.-H. (2014). Psychometrics behind computerized adaptive testing. *Psychometrika*, 80, 1-20.
- Chen, P., Xin, T., Wang, C., & Chang, H.-H. (2012). Online calibration methods for the DINA model with independent attributes in CD-CAT. *Psychometrika*, 77, 201-222.

- Guo, R., Zheng, Y., & Chang, H.-H. (2015). A stepwise test characteristic curve method to detect item parameter drift. *Journal of Educational Measurement*, 52, 280-300.
- Kingsbury, G. G. (2009). Adaptive item calibration: A process for estimating item parameters within a computerized adaptive test. In *The GMAC conference on computerized adaptive testing*. Retrieved from <http://iacat.org/sites/default/files/biblio/cat09kingsbury.pdf>
- Li, Y., & Lissitz, R. W. (2012). Exploring the full-information bifactor model in vertical scaling with construct shift. *Applied Psychological Measurement*, 36, 3-20. doi:10.1177/0146621611432864
- Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, 16, 159-176.
- Muraki, E. (1993). Information functions of the generalized partial credit model. *Applied Psychological Measurement*, 17, 351-363.
- Stocking, M. L. (1988). *Scale drift in online calibration* (Technical Report No. RR-88-28-ONR). Princeton, NJ: Educational Testing Service.
- Stocking, M. L. (1990). Specifying optimum examinees for item parameter estimation in item response theory. *Psychometrika*, 55, 461-475. doi:10.1007/BF02294761
- Tong, Y., & Kolen, M. J. (2007). Comparisons of methodologies and results in vertical scaling for educational achievement tests. *Applied Measurement in Education*, 20, 227-253.
- van der Linden, W. J., & Ren, H. (2015). Optimal Bayesian adaptive design for test-item calibration. *Psychometrika*, 80, 263-288.
- Veldkamp, B. P., Matteucci, M., & de Jong, M. G. (2013). Uncertainties in the item parameter estimates and robust automated test assembly. *Applied Psychological Measurement*, 37, 123-139. doi:10.1177/0146621612469825
- Wainer, H., & Mislevy, R. J. (2000). Item response theory, item calibration, and proficiency estimation. In H. Wainer (Ed.), *Computer adaptive testing: A primer* (pp. 65-102). Hillsdale, NJ: Lawrence Erlbaum.
- Wang, C., Chang, H.-H., & Boughton, K. A. (2013). Deriving stopping rules for multidimensional computerized adaptive testing. *Applied Psychological Measurement*, 37, 99-122. doi:10.1177/0146621612463422
- Zheng, Y. (2014). *New methods of online calibration for item bank replenishment* (Unpublished doctoral dissertation). University of Illinois at Urbana-Champaign.
- Zheng, Y., & Chang, H.-H. (2014, April). *The ordered informative range priority index (OIRPI) method for item selection in online calibration*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Philadelphia, PA.
- Zheng, Y., Wang, C., Culbertson, M., & Chang, H.-H. (2014, April). Overview of test assembly methods in multistage testing. In D. Yan, A. A. von Davier, & C. Lewis (Eds.), *Computerized multistage testing: Theory and applications* (pp. 87-99). New York, NY: CRC Press.