Listening. Learning. Leading.®

Research Report
ETS RR–14-38

An Item-Driven Adaptive Design for Calibrating Pretest Items

Usama S. Ali

Hua-Hua Chang

December 2014

ETS Research Report Series

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

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

ETS Research Report Series ISSN 2330-8516

RESEARCH REPORT

An Item-Driven Adaptive Design for Calibrating Pretest Items

Usama S. Ali[1] & Hua-Hua Chang[2]

1 Educational Testing Service, Princeton, NJ
2 University of Illinois at Urbana-Champaign, Champaign, IL

Adaptive testing is advantageous in that it provides more efficient ability estimates with fewer items than linear testing does. Item-driven adaptive pretesting may also offer similar advantages, and verification of such a hypothesis about item calibration was the main objective of this study. A suitability index (SI) was introduced to adaptively select pretest items, by which an easy-to-implement calibration methodology — adaptive design — can be used. A simulation study was conducted to evaluate the proposed adaptive design as compared to existing methodologies. Results indicate that the adaptive design has many desired features in item calibration, including less bias and more accurate parameter estimates, than the existing methods do. The SI is promising and flexible enough to apply additional constraints on the calibration sample and on the pretest items, for example, constraints on response time. It can also be used to try out individual item modules such as those used in multistage testing. Study limitations and future research are also covered in this report.

Keywords Adaptive design; computerized adaptive testing; item calibration; item response theory; pretest items; suitability index

doi:10.1002/ets2.12044

Computerized adaptive testing (CAT) has become more prevalent in large-scale assessment and has a glowing future in the K–12 context. CAT has succeeded in more quickly, accurately, and efficiently assessing abilities and skills of test takers for different purposes, such as licensure, placement, and college admission (Chang, 2012). Among many issues that are currently of concern for a large-scale CAT program, item pretesting is especially important. Administration of CATs requires precalibrating many new items, and therefore it is crucial to be able to calibrate test items in large quantities efficiently and economically (Stout, Ackerman, Bolt, Froelich, & Heck, 2003). Most testing programs tackle the problem by assembling new items into several blocks and assigning the blocks to test takers. These new items are pretested, by inserting them either individually in different parts of the test or collectively as a separate section of the test in a linear fashion (Tang & Eignor, 2001). As a result, the assignment of items is nonadaptive. According to Lord (1980), a test becomes the most effective when items are neither too difficult nor too easy to the test taker. Hence, if we assemble the blocks according to each test taker's estimated or partially estimated ability, such adaptive assignment may become more efficient than the traditional methods.

One challenge in item calibration of a CAT is to get accurate estimates at all trait levels. Because cut scores of major tests, such as the *ACCUPLACER*®, *Advanced Placement*® (*AP*®), and *GRE*® tests, to name a few, are determined by each institution, so there is no single score level that is critically important. Hence accuracy in estimating ability for each test taker is extremely important, and it is equally important at all different score levels. Thus getting enough high-quality items at each ability level is essential to increasing estimation efficiency for every test taker and, consequently, to increasing the test reliability. The reason is that item and test information is negatively influenced by errors in item parameter estimates (Gierl, Henderson, Jodoin, & Klinger, 2001; Hambleton & Jones, 1994). Therefore, accurately estimating parameters of items with different difficulty levels becomes essential as well.

Compared with other methods, CAT offers a better way of assigning additional items to each test taker for the sole purpose of calibrating new items. Because of the capability to allocate different items to different groups of individuals, the quality and quantity of test takers being exposed to a specific item can be controlled. One major advantage of CAT is that, compared with alternative methods, it provides more efficient ability estimates with fewer items. Using adaptivity may provide more accurate estimation of item parameters with fewer test takers than required in conventional paper-and-pencil tests or even within a CAT environment that uses an ordinary calibration design. From the CAT perspective,

Corresponding author: U. Ali, E-mail: uali@ets.org

adaptive item selection, therefore, provides an important potential application in calibrating pretest items. This may translate to fewer administrations of pretest items to obtain equally precise item parameter estimates as compared to calibrating without adaptive selection.

Few studies (e.g., Kingsbury, 2009; Makransky, 2009; Zhu, 2006) have attempted to use adaptivity to improve item calibration. The researchers in these studies investigated whether adaptive procedures could enhance the accuracy of item parameter estimation. The pretest item selection methods were similar to those used with operational items, for which more accurate ability estimation is the target (i.e., test-taker driven). Inspired by these leading studies, we explored whether a *multistage* item calibration methodology that involves *item-driven adaptive* selection of pretest items can help produce better item parameter estimates than those based on the traditional methods, such as preassignment or random assignment. The methodology is multistage in the sense of having multiple rounds of updated item parameter estimates as a result of increasing the calibration sample size. And it is adaptive in the method of exposing items to test takers. This methodology of calibrating pretest items is referred to as the *item-driven adaptive design* throughout this report.

In exploring the possible application of adaptivity to improving the quality of calibration samples for item parameter estimation with CAT, in this study, we had three specific objectives: (a) to provide evidence that improving the selection of the calibration sample through adaptive administration of pretest items is feasible in item parameter estimation, (b) to study a suitability index (SI) to adaptively select pretest items with an easy-to-implement calibration methodology, and (c) to evaluate the magnitude of the improvements in pretest item estimation for the CAT setting with the new methodology compared with existing methodology.

Item Calibration Within Computerized Adaptive Testing (CAT) Settings

Item calibration is an essential step in developing and implementing CATs. A central approach is the online item calibration, which refers to estimating the parameters of pretest items for future use during the course of current testing (Ban, Hanson, Wang, Yi, & Harris, 2001). Hence, the items administrated to a current test taker consist of two different types: operational items and pretest items. Actually, the two parts are carefully merged so that test takers cannot differentiate between them; otherwise, the quality of item pretesting will be affected. The need for online calibration has been increasing for several reasons: (a) the challenge of providing enough test takers for the calibration sample; (b) the differences in motivation between test takers in calibration samples, taking an isolated test of pretest items, relative to the potential population (which may result in biased item parameter estimates; Wise & DeMars, 2006); and (c) the influence of heavy use of web-based or online testing applications.

Item calibration in CATs involves different aspects, such as the calibration design, parameter estimation method, calibration sample, pretest item load, and seeding locations. There has been different emphasis on the various aspects of item pretesting. Most studies have focused on the estimation methods (e.g., Ban, Hanson, Yi, & Harris, 2002; Ban et al., 2001). Neglected in the literature have been other aspects of item pretesting, such as pretest design, including the item selection rules for pretesting, especially in a testing mode like CAT.

A major component of item calibration in a CAT environment is pretest design. Makransky (2009) developed three strategies for calibrating items and scoring test takers simultaneously using the same items. The strategies were two-phase, multiphase, and continuous updating strategies. It was assumed that there was no prior information available about items at the start of a new testing program. Two item response theory (IRT) models, one- and two-parameter logistic models, were considered. He concluded that these strategies performed very well compared with a fully calibrated test. These strategies were evaluated in terms of accuracy of ability estimation but were not evaluated directly in terms of item-related characteristics. Makransky's two-phase strategy is similar to Zhu's (2006) strategy. Initially, Zhu used a two-phase strategy to randomly select pretest items for a number of test takers. The first phase results in initial item parameter estimates. Then, at the second stage, these initial estimates are used to adaptively select pretest items for each test taker. This is similar to Kingsbury's (2009) study, in which he used the maximum information criterion to select pretest items. As is well known, item information is affected by two sources of error: item parameter estimates and ability parameter estimation. There are major challenges in using this criterion for selecting operational items; it is highly affected by the errors in item discrimination parameter estimates, especially for high a parameters, and has severely skewed item utilization (e.g., Chang, Qian, & Ying, 2001; Chang & Ying, 1999).

One practical decision in item calibration is when to stop sampling. There are usually two types of rules that are followed in making this decision: either a predetermined sample size has been reached or some standard error rule has been met.

Though a standard error rule may be more economical, it is more convenient to use a sample size rule (e.g., Wang & Wiley, 2004; Zhu, 2006). Ban et al. (2001) compared three sample sizes, 300, 1,000, and 3,000, while Ban et al. (2002) used 500 as a sample size, which is a reasonable choice for the marginal maximum likelihood (MML) method. Zhu (2006) chose three sample sizes for investigation: 300, 500, and 1,000. Note that Zhu used the two-parameter logistic model, whereas both Ban et al. (2001) and Ban et al. (2002) used the three-parameter logistic (3PL) model.

The quality of the calibration sample is also an important aspect. The characteristics of the calibration sample are crucial to achieving accurate item parameter estimates. Little emphasis has been directed toward the properties or qualities of a suitable sample for item calibration (e.g., Stocking, 1990). Even though a few studies used adaptive procedures to select pretest items, these procedures targeted the test-taker ability estimation (e.g., Makransky, 2009) or ignored the proper characteristics of a calibration sample for each item (e.g., Kingsbury, 2009). In this study, we emphasize using an appropriate sample to improve the accuracy of item parameter estimation. To implement the adaptive method of item calibration, a special group of test takers can be chosen, according to some criteria, from a particular large test-taker sample to calibrate each pretest item. The measurement model applied in this study is based on the 3PL IRT model.

Item-Driven Adaptive Design and Suitability Index (SI)

Previous research on adaptive administrations of pretest items focused on using these items for better scoring of test takers, not on improving calibration of the items. The focus was on the test takers in an effort to deliver the best possible items to them, while our focus here is on the item itself. The pretest item selection rule uses selection of the test takers to help enhance item parameter estimation. This is a major difference. We are proposing an adaptive design for online calibration of new items with a flexible design that allows for variation. The goal of the design is to obtain accurate item parameter estimates through enhancement of the quality of the calibration sample using a specific rule for selecting pretest items.

We propose in this report a simple strategy of adaptive assignment that uses item blocks or individual items (an individual pretest item is a special case of an item block) with the following steps:

1. Start administering these pretest items to a sample of test takers of size N_0.
2. Get initial item parameter estimates through IRT calibration results.
3. Group these items into a number of blocks (e.g., five blocks) that differ in average difficulty using the results from Step 2.
4. For each test taker, adaptively choose one block or individual items either through matching the ability estimate to the difficulty estimate or through another appropriate criterion for selecting pretest items (as will be explained later in this section), and administer the block.
5. Update the item parameter estimates based on data gained from Step 4.
6. Repeat Steps 4 and 5 until stable item parameter estimates (e.g., minimum parameter estimate change) are obtained, or satisfy another stopping rule (e.g., sample size).

The proposed method is refined by changing the sample size and the number of item blocks within each round. Another possible dimension in item block formation is item content. Blocks may be assembled such that item content and item types will be approximately balanced. Because each block consists of only two to four items, there is no way to balance content strictly. Nonetheless, content can be balanced in the sense that no test taker receives only items of the same content.

The basic idea of adaptive design for testing new items is the same as that for adaptively selecting operational items for estimating test takers' abilities. The design is flexible to the extent that it may be completely adaptive, where individual items are selected and administered, one at a time. In this case, each block in Step 3 comprises only one item. An alternative is to have several blocks, each comprising several items suitable for a given ability range, which corresponds in a sense to multistage testing.

Stocking (1990) provided some guidelines for choosing a sample of specific qualities in testing situations such as calibrating item parameters. These guidelines assist in accurately estimating item parameters for a collection of items for which the properties are not known for certain. Stocking concluded that, if the 3PL model is considered for item analysis, (a) the optimal and most informative sample of test takers for best estimation of item slope is a combination of some with low ability levels and some with high ability levels; (b) the optimal and most informative sample of test takers for best estimation of item location of easy and difficult items is one within the neighborhood of the item difficulty parameter; and (c) the optimal and most informative sample for best estimation of item lower asymptote is one of low ability. For estimating

all item parameters, a wide distribution of abilities (e.g., uniform) is more informative than a bell-shaped distribution. Stocking selected those optimum levels of the latent trait for item calibration using Fisher's information matrix, as used in optimal designs. In D-optimal designs, the objective function is the determinant of the information matrix or its inverse for calibrating dichotomous items (Berger, King, & Wong, 2000) and polytomous items (Holman & Berger, 2001).

To select the most suitable pretest item, a pretest item SI may be used. This procedure of using such an index takes into consideration the requirements that need to be met for accurate item calibration and results in the selection of a suitable item or set of items for each test taker. For each item, the calibration sample should satisfy some constraints, such as the range of ability estimates or minimum sample size.

Let us denote a sample requirement or tracing matrix by C, where C is a J by K matrix. Each entry c_{jk} in the matrix provides the number of test takers in a specific ability range k, $k = 1,2, \ldots, K$, who responded to a specific item j, $j = 1,2, \ldots, J$. The SI for item j, S_j is calculated by:

$$S_j = \frac{1}{\left|\widehat{b}_j - \widehat{\theta}\right|} \prod_{k=1}^{K} w_k f_{jk}, \quad j = 1, 2, \cdots, J, \tag{1}$$

where $\widehat{b}_j$ is the difficulty parameter estimate of the pretest item j, $\widehat{\theta}$ is the current ability estimate, w_k is the weight assigned to ability range k, $k = 1,2, \ldots, K$, and f_{jk} is the proportion of test takers from a specific range, k, that is still needed to meet the target. Then this proportion is given by:

$$f_{jk} = \frac{T_{jk} - t_{jk}}{T_{jk}}, \tag{2}$$

where T_{jk} is the target sample size from ability range k for item j and t_{jk} is the number of test takers who already responded to the item. The values of T_{jk} can be determined based on the initial item parameter estimates resulting from the random assignment stage. By maximizing the S_j, specifications for the calibration sample will be fulfilled.

We use this index to ensure exposure of pretest items to test takers with needed characteristics within the practical boundaries of an operational program (i.e., the ability distribution of the population). In other words, the goal is to have a better calibration sample to estimate the statistical properties of an item.

Therefore, the selection and administration of pretest items during the operational test administration is based on a different procedure than that used to select operational items. The pretest item selection procedure is used to choose items for which the current test taker is considered optimum with respect to item parameter estimation.

Data and Computerized Adaptive Testing (CAT) Simulation

In this study, we used Monte Carlo simulation to address its objectives. This section includes data description of the real operational items and simulated pretest items and details on the CAT simulation.

Data

In this study, we used data from a large-scale operational test in elementary algebra. The elementary algebra test is an adaptive test assessing mathematics skills. The item parameters for all items were estimated assuming a 3PL item response model. The operational and pretest items are described in the following sections.

Operational Items

The operational item pool has 521 active items. The elementary algebra test is a short-length test consisting of 12 items chosen to meet certain constraints. Constraints take the form of overlap (O) and content (C) codes (see Table 1). Items with the same overlap code cannot appear in the same test; they are considered enemy items. The content code specifies the content topic or topics that an item covers. In total, there are 15 constraints of type O and 21 constraints of type C. Content constraints specify three main topics: (a) signed numbers and rationales (C1 – C2), (b) algebraic expressions (C3 – C10), and (c) equations, inequalities, and word problems (C11 – C15). Other codes specify inequalities (C17), negative stem (C16), and key distribution (C18 – C21).

Table 1 Distribution of Items in Different Content Areas and Constraints Codes

Topic	Content code	No. of items	Weight	Minimum	Maximum
Signed numbers and rationales	C1	20	5	0	1
	C2	45	15	1	1
Algebraic expressions	C3	23	10	0	1
	C4	22	10	0	1
	C5	53	20	1	1
	C6	33	15	1	1
	C7	44	15	1	1
	C8	22	5	0	1
	C9	32	15	1	1
	C10	22	15	1	1
Equations, inequalities, and word problems	C11	69	15	1	2
	C12	18	5	0	1
	C13	73	5	0	1
	C14	25	10	1	1
	C15	20	5	0	1
Negative stem	C16	7	5	0	1
Inequalities	C17	16	5	0	2
Key	C18	134	0.5	2	5
	C19	130	0.5	2	5
	C20	131	0.5	2	5
	C21	126	0.5	2	5

Table 2 Descriptive Statistics of the Operational Item Pool

Parameter	*N*	Mean	*SD*	Minimum	Maximum
a	521	1.163	0.485	0.293	2.926
b	521	0.042	1.097	− 4.790	3.624
c	521	0.199	0.092	0	0.500

Simple descriptive statistics (i.e., mean, standard deviation, minimum, and maximum) of the operational item pool are provided in Table 2. Depicting item pool information illustrates the potential test-taker distribution that can be optimally tested by the operational item pool. This helped simulate the ability distribution. As illustrated in Figure 1, the item pool covers a relatively wide θ range. The pool information curve is nearly symmetric and reaches its peak around $\theta = 0.6$.

Pretest Items and Calibration Method

The pretest item pool has the same characteristics as the operational item pool. Fifteen pretest items were selected for the current research. Pretest item parameters were generated to match the characteristics of the operational item pool (see Table 3). These items approximately range in difficulty level from − 3.0 to 3.0 and in discrimination power from 0 to 2.0.

Regarding the calibration sample, we chose six sample sizes for study: 300, 500, 750, 1,000, 1,500, and 2,000. That is, $N_0 = 300$ is for the first stage, where the items are randomly selected for each test taker. From this stage, initial item parameter estimates were obtained using Parscale (Muraki & Bock, 2003). Individual pretest items were selected and delivered, one at a time (i.e., item block size = 1). Then, we updated the item parameter estimates after we reached the specified sample sizes.

According to Ban et al. (2001), the multiple-cycle expectation-maximization procedure is the best choice for applying MML estimation, which was used in this study. This procedure requires two important configurations: the item parameters of the operational items are fixed at their original values, and two or more EM cycles are used to estimate item parameters of the pretest items.

To obtain the item parameter estimates, some specifications were chosen: The maximal number of EM cycles was set to 100, the number of Gauss–Newton iterations was 10, and the convergence criterion for EM and Newton iterations

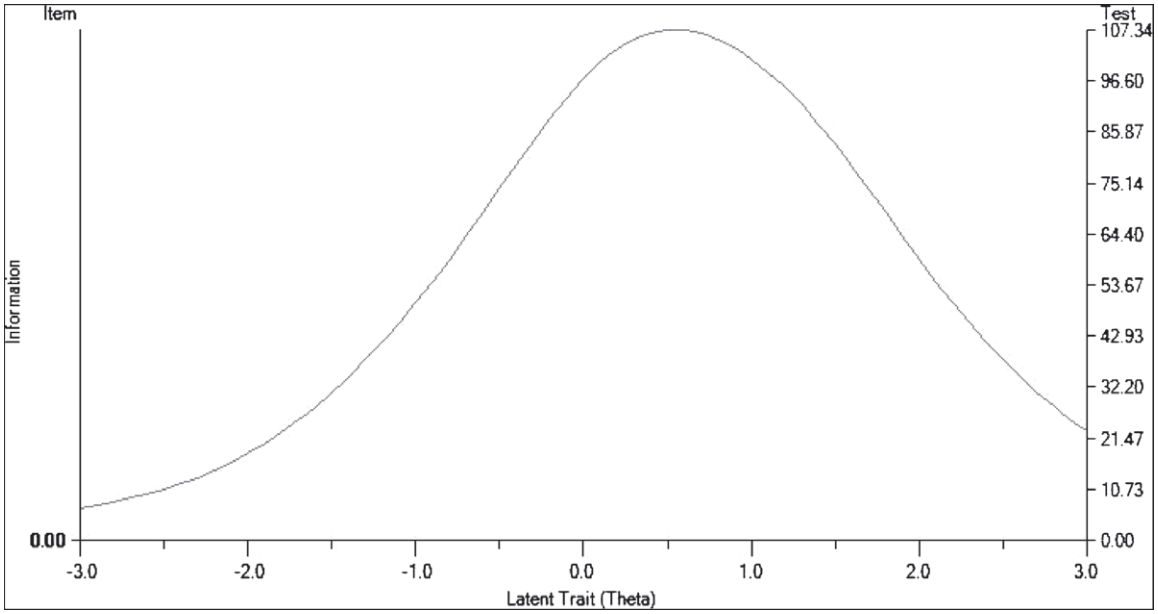

Figure 1 Total item pool information.

Table 3 Parameters of Pretest Items

Item	*a*	*b*	*c*
1	1.293	− 0.862	0.012
2	0.912	2.856	0.210
3	1.544	2.285	0.118
4	0.676	1.792	0.109
5	0.833	− 0.295	0.097
6	0.927	− 2.952	0.238
7	1.131	0.745	0.245
8	0.609	− 1.651	0.255
9	0.571	0.023	0.038
10	0.334	− 2.493	0.079
11	1.824	1.168	0.128
12	0.715	− 1.205	0.168
13	1.125	0.537	0.194
14	0.513	− 2.121	0.121
15	1.602	1.383	0.225

was set to 0.001. Using these configurations, no convergence problems were reported in any estimation session, and the estimation met the convergence criteria in fewer than 20 cycles.

CAT Simulation

Ability Distribution

The calibration sample of 10,000 simulees was generated from a standard normal distribution, $N(0, 1)$, to suitably match the item pool (see Figure 1). This volume of test takers is considered low relative to the average flow of test takers taking the test every month. The sample size is an important factor in determining the adequacy of the adaptive strategy of pretest item selection. The 10,000 generated θs are considered true θs. The test takers were classified into eight categories (i.e., $\theta < -3, -3 \leq \theta < -2, \ \ldots, 2 \leq \theta < 3$, or $\theta \geq 3$). This categorization was used to track the coverage of ability range of the test takers exposed to each pretest item.

Test Specification

A small number of pretest items in operational CATs were seeded through the operational test. As a rule of thumb, the seeded items did not exceed 25–33% of the test length. Therefore, owing to the short length of elementary algebra tests (i.e., 12 items), three pretest items (i.e., representing 25% of the test) were chosen to be added to each test.

Hence a testing session consisted of 12 operational items and 3 pretest items, which delivers a 15-item test in total to each test taker. The content constraints for the operational items are reported in Table 1. In this study, the pretest items were administered for each test taker after administering all the 12 operational items.

Item Selection Method

The operational item selection method considered for this research is the maximum priority index (MPI; Cheng & Chang, 2009; Cheng, Chang, Douglas, & Guo, 2009). The MPI was selected because of the specifications of the elementary algebra test, which constitutes a severely constrained test. The MPI performs very well in handling the constraints with minimal violations. The constraints are displayed in Table 1. In addition to such content constraints, we also added item exposure control as an additional constraint for test security. The upper bound of exposure rate was set at .30 in the current analysis. This method was used to select the 12 operational items per test taker.

For the pretest item selection, three item selection methods were considered for comparison: the maximum suitability index (MSI), the minimum difference between the current ability estimate of a test taker and item difficulty (match-b), and the random selection, a baseline measure (random).

Scoring Method

The expected a posteriori method was used to estimate a test taker's ability for at least the first five items, and then the maximum likelihood estimation was used if the response pattern allowed (response patterns comprising all zeros or all ones do not yield maximum likelihood estimates).

Evaluation Criteria

Item parameter recovery was used to evaluate the proposed adaptive pretesting method as compared to traditional methods. Also, the sample size to achieve accurate item parameter estimation was considered, as we assumed we would have smaller sample sizes compared with nonadaptive methods. Evaluation criteria that we used were (a) average bias and (b) root mean square difference (RMSD).

Both indices capture measurement precision. Average bias is estimated using Equation 3. In Equation 3, let β_i, $i = 1, \ldots, n$ be the true item parameters of n items and $\widehat{\beta}_i$ be the respective estimators using different calibration designs of different pretest item selections rules.

Then the estimated average bias is computed as:

$$\text{Bias} = \frac{1}{n}\sum_{i=1}^{n}\left(\beta_i - \widehat{\beta}_i\right). \tag{3}$$

RMSD is calculated using Equation 4, as follows:

$$\text{RMSD} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\beta_i - \widehat{\beta}_i\right)^2}. \tag{4}$$

The smaller the average bias and RMSD, the better the calibration strategy used.

Results

Figures 2 and 3 depict the effect of calibration sample size on the estimation accuracy of a, b, and c in terms of bias and RMSD for all the studied methods. All methods tended to overestimate item parameters, as indicated by the negative biases in Figure 2. On average, the biases in item parameter estimates are relatively small as a percentage of the possible ranges

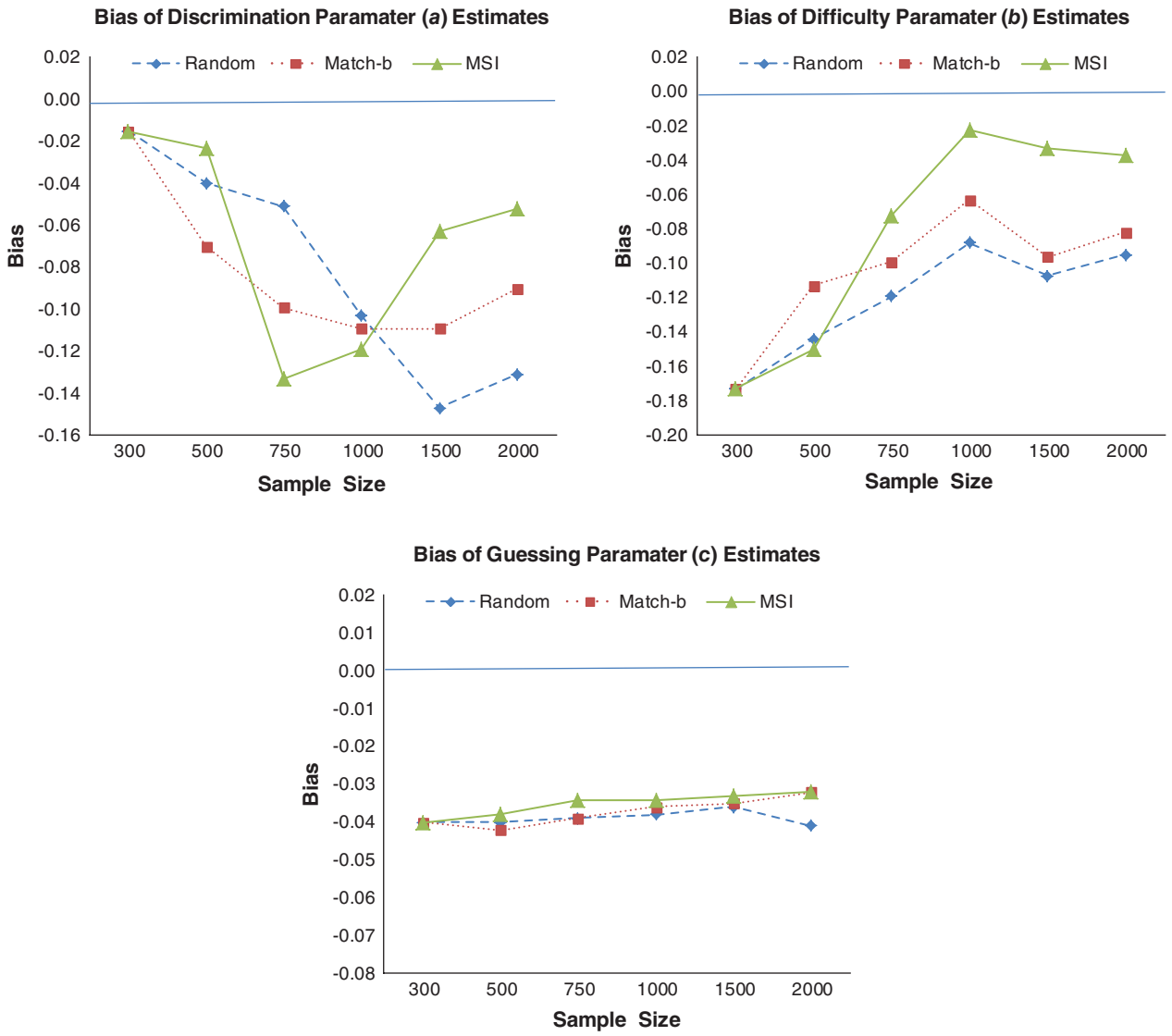

Figure 2 Bias of item parameter estimates.

of these parameters (e.g., b values fall within $[-3.0, 3.0]$). The mean biases in estimating a and b parameters ranged from -0.015 to -0.147 and from -0.022 to -0.173, respectively, while biases ranged from -0.032 to -0.042 for estimating the c parameter.

Given the calibration sample size, the MSI consistently provided lower RMSDs compared with the match-b and random methods for a and b parameters. For the b parameter, the MSI method with a sample size of 500 provided an RMSD value that is equivalent to that attained by the match-b method with a sample of 1,500 test takers. It is even more accurate than was attained by the random selection method with a sample size of 1,000. All methods provided comparable results for the c parameter. In general, the adaptive methods (i.e., both the MSI and match-b methods) provided better results than the random method.

With regard to the performance of each method, there exist some irregularities. It is assumed that by increasing the calibration sample size, the sample becomes more representative of the population, and the variation in item parameter estimates tends to decrease. In Figure 3 (top), the RMSD for the match-b method increased from 0.209 to 0.263 as the sample size increased from 750 to 1,000—contrary to this assumption. By investigating the actual estimates of a, we found that Item 2, $(a_2, b_2, c_2) = (0.912, 2.856, 0.210)$, is poorly estimated, $\left(\hat{a}_2, \hat{b}_2, \hat{c}_2\right) = (1.773, 2.486, 0.234)$. The large difference between the true and estimated a parameter, 0.861, explains this jump in the RMSD. Note that Item 2 is a difficult

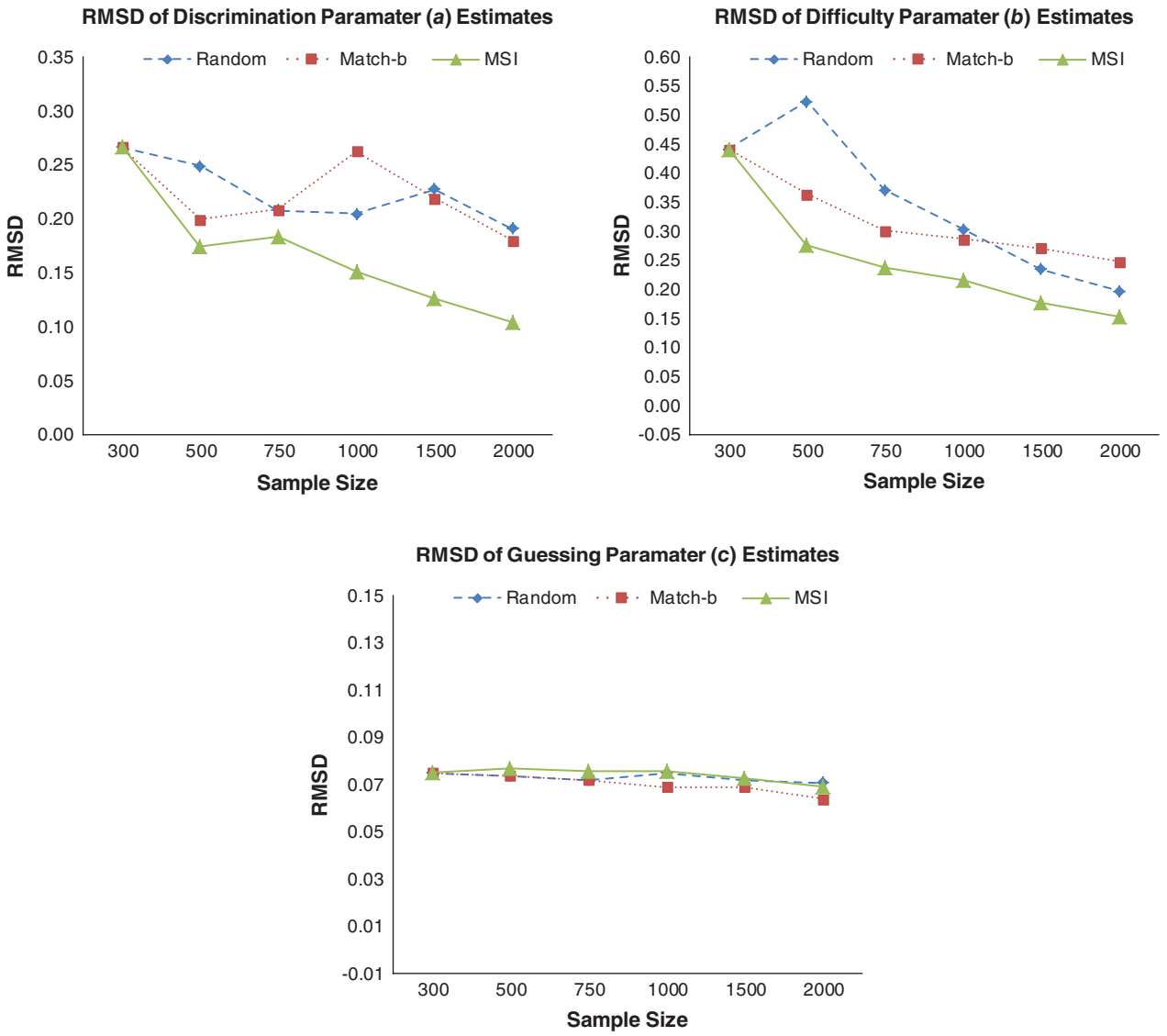

Figure 3 RMSD of item parameter estimates.

item with medium discrimination. In Figure 3 (middle), an irregular result exists for the random method. The RMSD increased from 0.442 at a sample size of 300 to 0.524 at a sample size of 500. Again, we found two poorly estimated items, Items 2 and 6. In the 500 sample size case, $\left(\hat{a}_2, \hat{b}_2, \hat{c}_2\right) = (0.656, 4.034, 0.242)$ and $\left(\hat{a}_6, \hat{b}_6, \hat{c}_6\right) = (0.571, -4.260, 0.205)$, with differences in b parameters of -1.178 and 1.308, respectively. Conversely, Item 6 is representative of the easiest items.

To evaluate the different strategies for pretesting items, we look at the accuracy of item calibration results against the true item parameters. Tables 4 and 5 provide data on the overall estimation accuracy of the 3PL item parameters: a, b, and c. The mean, standard deviation, minimum and maximum bias, and RMSD throughout the different runs are presented. It is obvious that the MSI method consistently provided lower biases and RMSDs in estimating item parameters a, b, and c.

Discussion

As is well known, an advantage of CAT is that it provides more efficient ability estimation with fewer items than other methods. It has also been assumed that adaptive pretesting strategies provide more precise item parameter estimates with fewer test takers than is required in nonadaptive procedures such as random assignment. Verification of such a hypothesis about item calibration was the main objective of this study.

Table 4 Summary Statistics of Bias of 3PL Item Parameter Estimates

Method	Mean	*SD*	Minimum	Maximum
Parameter *a*				
Random	− 0.081	0.053	− 0.015	− 0.147
Match-*b*	− 0.082	0.036	− 0.015	− 0.109
MSI	−0.067	0.049	− 0.015	− 0.133
Parameter *b*				
Random	− 0.121	0.032	− 0.088	− 0.173
Match-*b*	− 0.104	0.038	− 0.063	− 0.173
MSI	−0.081	0.065	− 0.022	− 0.173
Parameter *c*				
Random	− 0.039	0.002	− 0.036	− 0.041
Match-*b*	− 0.037	0.003	− 0.032	− 0.042
MSI	−0.035	0.003	− 0.032	− 0.040

Table 5 Summary Statistics of RMSD of 3PL Item Parameter Estimates

Method	Mean	*SD*	Minimum	Maximum
Parameter *a*				
Random	0.224	0.029	0.191	0.267
Match-*b*	0.223	0.035	0.180	0.267
MSI	0.168	0.057	0.104	0.267
Parameter *b*				
Random	0.346	0.125	0.198	0.524
Match-*b*	0.319	0.072	0.249	0.442
MSI	0.251	0.104	0.154	0.442
Parameter *c*				
Random	0.073	0.002	0.071	0.075
Match-*b*	0.070	0.004	0.064	0.075
MSI	0.074	0.003	0.069	0.077

The results provided evidence that an item-driven adaptive design for item pretesting has many desired features in item calibration: less bias and more accurate parameter estimation. These results were obtained using small sample sizes. Although the sample size is not considered to be an issue for the studied test, producing accurate parameter estimates of new items with relatively small sample sizes is desirable and needed for other tests that have lower volumes of test takers. The focus on increasing the accuracy of item parameter estimation would actually help enhance the test development process and avoid the problems resulting from less accurate or inaccurate parameter estimates (Hambleton, Jones, & Rogers, 1993).

The selection of test takers with a wide ability range can help obtain a relatively representative sample and avoid bias in item parameter estimates. It is clear that restricting the range of ability affects the results of the pretest calibration. This is illustrated by the poor parameter estimation of the most difficult and easiest items at the small sample sizes, especially with the random and match-*b* strategies. In these cases, the MSI performed better by producing a more suitable sample. The adaptive and random selection strategies followed in this study had the following limitations: a small number of pretest items (three items were administered to each test taker), the need to use equal sample sizes, and minimum exposure to a sufficient number of test takers with ability levels along the entire range of ability. Wang and Wiley (2004) reported that the general rule of thumb is to use about 800 test takers per item as a sound sample size. But given all these circumstances, the adaptive strategy using the MSI method for selecting pretest items was found to be superior to the other studied methods.

Considering that the operational test is not speeded, we chose to deliver pretest items as a separate section at the end of the test. Generally, measurement errors of ability estimates are expected to be large at the early stage of CATs (van der Linden & Pashley, 2000). Hence, we recommend that pretest items be placed later in the test to reduce the effect of poor ability estimation on the selection of items. Thus pretest items would be placed around the final stage of an operational test, where more reliable ability estimates emerge. For example, for a 12-item test, three pretest items might be placed after administering the 8th, 10th, and 12th items.

To enhance the performance of the adaptive design, the MSI method is promising and flexible enough to apply any set of desired constraints on the calibration sample. For example, we can put a constraint regarding the total response time needed to answer all pretest items given to an individual test taker. The mean (or median) response time of a pretest item can be obtained from the initial stage of administering such an item. Estimates of the response time are generally useful in practice for balancing the test load among test takers. This would apply not only to operational items but also to pretest items. The MSI method should be evaluated by comparing it to other methods of administering pretest items, such as using a prespecified criterion of item parameter change.

The generalizability of the current report's results is limited to dichotomous items fitted by 3PL models, and test takers' abilities were estimated from a 12-item test. Only 15 pretest items were considered; a larger pool of such items may provide a better variety of items for future use on operational tests. The validity of the results should be verified for dichotomous items analyzed by other IRT models (e.g., 1PL and 2PL) and for polytomous items as well. As known, adaptive testing enhances ability estimation more for extreme test takers. Also, we could investigate the extent to which the adaptive selection of the calibration sample helps the parameter estimation of extreme-difficulty or high- or low-discrimination items. In this study, we applied the proposed adaptive design at the item level. The design is suitable; it can be used with item blocks or item modules that are administered together (see Ali & Liang, 2013). Such flexibility allows using the adaptive method in multistage testing, where one item module is administered at a time. Finally, the adaptive method could be modified to include other constraints in the calibration samples, and it could be compared to optimal designs in item parameter estimation (e.g., see Berger et al., 2000).

References

Ali, U. S.., & Liang, L. (2013, July). *Item-driven adaptive design of pretesting items in multistage testing*. Paper presented at the 12th international and the 78th annual meeting of the Psychometric Society, Amsterdam, the Netherlands.

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.

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.

Berger, M. P. F., King, C. Y. J., & Wong, W. K. (2000). Minimax D-optimal designs for item response theory models. *Psychometrika*, *65*, 377 – 390.

Chang, H.-H. (2012). Making computerized adaptive testing diagnostic tools for schools. In R. W. Lissitz, & H. Jiao (Eds.), *Computers and their impact on state assessment: Recent history and predictions for the future* (pp. 195 – 226). Charlotte, NC: Information Age.

Chang, H.-H., Qian, J., & Ying, Z. (2001). *a*-stratified multistage computerized adaptive testing with *b* blocking. *Applied Psychological Measurement*, *25*, 333 – 341.

Chang, H.-H., & Ying, Z. (1999). *a*-stratified multistage computerized adaptive testing. *Applied Psychological Measurement*, *23*, 211 – 222.

Cheng, Y., & Chang, H.-H. (2009). The maximum priority index method for severely constrained item selection in computerized adaptive testing. *British Journal of Mathematical and Statistical Psychology*, *62*, 369 – 383.

Cheng, Y., Chang, H.-H., Douglas, J., & Guo, F. (2009). Constraint-weighted *a*-stratification for computerized adaptive testing with nonstatistical constraints. *Educational and Psychological Measurement*, *69*, 35 – 49.

Gierl, M. J., Henderson, D., Jodoin, M., & Klinger, D. (2001). Minimizing the influence of item parameter estimation errors in test development: A comparison of three selection procedures. *The Journal of Experimental Education*, *69*, 261 – 279.

Hambleton, R. K., & Jones, R. W. (1994). Item parameter estimation errors and their influence on test information functions. *Applied Measurement in Education*, *7*, 171 – 186.

Hambleton, R. K., Jones, R. W., & Rogers, H. J. (1993). Influence of item parameter estimation errors in test development. *Journal of Educational Measurement*, *30*, 143 – 155.

Holman, R., & Berger, M. P. F. (2001). Optimal calibration designs for tests of polytomously scored items described by item response theory models. *Journal of Educational and Behavioral Statistics*, *26*, 361 – 380.

Kingsbury, G. G. (2009). Adaptive item calibration: A process for estimating item parameters within a computerized adaptive test. In D. J. Weiss (Ed.), *Proceedings of the 2009 GMAC conference on computerized adaptive testing*. Retrieved from http://www.psych.umn.edu/psylabs/CATCentral/

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Erlbaum.

Makransky, G. (2009). An automated online calibration design in adaptive testing. In D. J. Weiss (Ed.), *Proceedings of the 2009 GMAC conference on computerized adaptive testing*. Retrieved from http://www.psych.umn.edu/psylabs/CATCentral/

Muraki, E., & Bock, R. D. (2003). *PARSCALE 4: IRT based test scoring and item analysis for graded items and rating scales* [Computer software]. Chicago, IL: Scientific Software International.

Stocking, M. L. (1990). Specifying optimum examinees for item parameter estimation in item response theory. *Psychometrika*, *55*, 461–475.

Stout, W., Ackerman, T., Bolt, D., Froelich, A. G., & Heck, D. (2003). *On the use of collateral item response information to improve pretest item calibration* (Computerized Testing Report No. 98–13). Newtown, PA: Law School Admission Council.

Tang, K. L., & Eignor, D. R. (2001). *A study of the use of collateral statistical information in attempting to reduce TOEFL IRT item parameter estimation sample sizes* (Technical Report No. TR-17). Princeton, NJ: Educational Testing Service.

van der Linden, W. J., & Pashley, P. J. (2000). Item selection and ability estimation in adaptive testing. In W. J. van der Linden, & C. A. W. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (pp. 1–25). Boston, MA: Kluwer.

Wang, X. B., & Wiley, A. (2004, April). *Achieving accuracy of pretest calibration for a national CAT placement examination with a restricted test length*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego, CA.

Wise, S. L., & DeMars, C. E. (2006). Low examinee effort in low-stakes assessment: Problems and potential solutions. *Educational Assessment*, *10*, 1–17.

Zhu, R. (2006). *Implementation of optimal design for item calibration in computerized adaptive testing (CAT)* (Unpublished doctoral dissertation). University of Illinois, Urbana.

Suggested citation:

Ali, U. S., & Chang, H.-H. (2014). *An item-driven adaptive design for calibrating pretest items* (Research Report No. RR-14-38). Princeton, NJ: Educational Testing Service. doi:10.1002/ets2.12044

Action Editor: James Carlson

Reviewers: Lixiong Gu and Deping Li