ACT Research Report Series 2000-11

A Comparative Study of Online Pretest Item Calibration/ Scaling Methods in CAT

Jae-Chun Ban

Bradley A. Hanson

Tianyou Wang

Qing Yi

Deborah J. Harris

ACT

December 2000

For additional copies write:
ACT Research Report Series
PO Box 168
Iowa City, Iowa 52243-0168

A Comparative Study of Online Pretest Item Calibration/Scaling Methods in CAT

Jae-Chun Ban
Bradley A. Hanson
Tianyou Wang
Qing Yi
Deborah J. Harris

Abstract

The purpose of this study was to compare and evaluate five online pretest item calibration/scaling methods in computerized adaptive testing (CAT): the marginal maximum likelihood estimate with one EM cycle (OEM) method, the marginal maximum likelihood estimate with multiple EM cycles (MEM) method, Stocking's Method A, Stocking's Method B, and the BILOG/Prior method. The five methods were evaluated in terms of item parameter recovery under three different sample size conditions (300, 1,000, and 3,000). The MEM method appears to be the best choice among the methods used in this study, because it produced the smallest parameter estimation errors for all sample size conditions. Stocking's Method B also worked very well, but it requires anchor items, which would make test lengths longer. The BILOG/Prior method did not seem to work with small sample sizes. Until more appropriate ways of handling the sparse data with BILOG are devised, the BILOG/Prior method may not be a reasonable choice. Because Stocking's Method A has the largest weighted total error, as well as a theoretical weakness (i.e., treating estimated ability as true ability), there appears to be little reason to use it. The MEM method should be preferred to the OEM method unless amount of time involved in iterative computation is a great concern. Otherwise, the OEM method and the MEM method are mathematically similar, and the OEM method produces larger errors than the MEM method.

A Comparative Study of Online Pretest Item Calibration/Scaling Methods in CAT

Introduction

In computerized adaptive testing (CAT), pool replenishing is a necessary process for maintaining an item pool because items in the pool would be obsolete or overexposed as time goes on. To be added as new items in the pool, the pretest items should be calibrated and be on the same scale as items already in the pool.

Online calibration refers to estimating the parameters of pretest items which are presented to examinees during the course of their testing with operational items (Stocking, 1988; Wainer & Mislevy, 1990). Since the item parameter estimates obtained from the paper and pencil delivery mode are not necessarily comparable to the item parameter estimates calibrated from the CAT mode, due to such factors as item ordering, different mode of test administration, context, local item dependency, and cognitive difference (Parshall, 1998; Spray, Parshall, & Huang, 1997), pretest item calibration/scaling methods that utilize the online testing system should be developed.

The complication in online pretest item calibration results from the fact that, typically, item response data obtained from CAT administrations are sparse, based on a restricted range of ability (Folk & Golub-Smith, 1996; Haynie & Way, 1995; Hsu, Thompson, & Chen, 1998; Stocking, 1988), and a relatively small number of items are administered compared to the paper and pencil delivery mode. These data characteristics may lead to inaccurate item parameter estimates for the pretest items. Nevertheless, the online pretest item calibration/scaling has several advantages such as preserving testing mode and utilizing the pretest data obtained during operational testing, and reducing the impact of motivation and representativeness concerns coming from the administration of pretest items to volunteers (Parshall, 1998).

Several studies have proposed online pretest item calibration methods (Folk & Golub-Smith, 1996; Levine & Willams, 1998; Samejima, 2000; Stocking, 1988; Wainer & Mislevy, 1990). Among them, some methods involve using parametric item response functions in which pretest item characteristic curves are estimated as a three-parameter logistic model, whereas other methods employ nonparametric methods of estimating item response functions. In the present study, pretest item calibration/scaling methods that use the parametric item response model were compared.

Although it is valuable to identify the general properties of each method, it is difficult to compare and evaluate results across studies because most studies have included only one or two methods, and used different research designs and criteria.

It is necessary that these online pretest calibration/scaling methods be compared under identical conditions to reveal their relative strengths and weaknesses in pretest item parameter estimation. The purpose of this study was to compare and evaluate parametric online pretest item calibration/scaling methods in terms of item parameter recovery for different sample sizes. In the next section, we discuss the characteristics of each online calibration method in greater detail.

Online Pretest Item Calibration Methods

One-EM Cycle Method

Wainer and Mislevy (1990, pp. 90-91) described the marginal maximum likelihood estimate with one EM cycle (OEM) approach for calibrating online pretest items. In this paper, the three-parameter (3-PL) logistic item response model is used to model item responses. For the 3-PL model, the probability of a correct response to item i by an examinee with ability θ_j is

$$P(u_{ij} = 1 \mid \theta_j) = P_i(\theta_j) = c_i + \frac{1 - c_i}{1 + e^{-Da_i(\theta_j - b_i)}} \qquad (1)$$

where a_i, b_i, and c_i are the item discrimination, difficulty parameter, and lower asymptote of item i (i=1,..., n), respectively, and D is the constant 1.7. The likelihood function of observing an item response, u_{ij}, on an operational item for an examinee with ability θ_j (j=1,..., N) is

$$L(u_{ij} \mid \theta_j) = P_i(\theta_j)^{u_{ij}}[1 - P_i(\theta_j)]^{1-u_{ij}}. \tag{2}$$

Similarly, the likelihood of observing the response, x_{kj} (k=1,..., K), on a pretest item for an examinee with ability θ_j is

$$L(x_{kj} \mid \theta_j) = P_k(\theta_j)^{x_{kj}}[1 - P_k(\theta_j)]^{1-x_{kj}}. \tag{3}$$

The joint likelihood of N different item responses on a pretest item is the product of the separate likelihoods. This joint likelihood is

$$L = \prod_{j=1}^{N} L(x_{kj} \mid \theta_j) = \prod_{j=1}^{N} P_k(\theta_j)^{x_{kj}}[1 - P_k(\theta_j)]^{1-x_{kj}}. \tag{4}$$

In the OEM approach, the item parameters are estimated by maximizing the marginal likelihood. OEM takes just one E-step using the posterior distribution of ability, which is estimated based on item responses only from the operational items, and just one M-step to estimate item parameters, involving data from only the pretest items (Wainer & Mislevy, 1990). The M-step finds parameter estimates that maximize

$$L^* = \prod_{j=1}^{N} \int L(x_{kj} \mid \theta) g(\theta \mid u_j, \beta_{operational}) d\theta, \tag{5}$$

where $g(\theta)$ is a posterior distribution of θ given the responses to the operational items and item parameters for the operational items, and $\beta_{operational}$ is a vector of the known item parameters of the operational items. Maximization of Equation 5 with respect to item parameters produces parameter estimates of the pretest items based on one M-step in the EM algorithm for marginal maximum likelihood estimates. With this approach, the item parameter estimates of the pretest items would only be updated once because only one M-step of the EM cycle is computed.

In implementing this method in practice, a Bayesian modal estimation approach may be used by multiplying the marginal maximum likelihood equations by a prior distribution for the pretest item parameters.

Some advantages of this approach are that since the operational items are used to compute the posterior ability distribution, the pretest items are automatically on the same scale as the operational items, and that no pretest item can contaminate other pretest items because the pretest items are calibrated independently of other pretest items (Parshall, 1998).

Multiple-EM Cycles Method

As a variation of the OEM method, we increased the number of EM cycles until convergence criterion was met. This method is called here the marginal maximum likelihood estimate with multiple EM cycles (MEM) method. MEM is very similar mathematically to OEM. The first EM cycle with MEM is the same as OEM approach. That is, MEM computes the posterior distribution using the operational items and finds item parameters that maximizes Equation 5 to obtain item parameter estimates.

However, beginning with the second E-step, MEM uses item responses on both the operational items and pretest items to get the posterior distribution. It should be noted that for each M-step iteration, the item parameter estimates for the operational items are fixed, whereas parameter estimates for the pretest items are updated until the item parameter estimates converge. With this method, the pretest items are also automatically on the same scale as items in the pool.

A prior distribution for pretest item parameters may be assumed when the MEM method is implemented in practice, in which case the resulting parameter estimates are Bayes modal estimates.

One important advantage of this method is that it fully uses information from item responses on pretest items for calibration by taking multiple EM cycles. Because this method uses the item responses on the pretest items in the E-step, however, some poor pretest items may affect the computation of the posterior distribution from the second E-step and the resulting pretest item calibrations, particularly when the number of operational items is small (e.g. 10 items).

BILOG with Strong Prior Method

This method uses the computer program BILOG with strong priors on the operational items. By putting strong priors on the operational items, the BILOG with Strong Prior (BILOG/Prior) method in essence fixes the operational item parameters while estimating pretest item parameters. As an example, Folk and Golub-Smith (1996) calibrated the operational items concurrently with the pretest items and anchor items using BILOG. They used an option in BILOG to put strong priors on the operational, pretest, and anchor items. The original item parameter estimates were specified as means of strong priors for the operational CAT items and the means of item parameter estimates in the CAT item pool were designated as priors for pretest and anchor items.

We used different priors in this study from those used in Folk and Golub-Smith (1996). We put strong priors on the item parameter estimates for operational items by setting the prior means equal to the their calibrated parameter estimates with very small prior variances. Default priors were used for the item parameter estimates for pretest items. More specific procedures are described in the methods section. When using the BILOG/Prior method, however, the re-estimated operational item parameter estimates would be different from the operational item parameter estimates that were in the item pool, depending on the magnitude of the prior standard deviations for each parameter. Note that MEM does not re-estimate the operational item parameter estimates, but instead uses the previously obtained item parameter estimates.

BILOG/Prior and MEM are the same in that both methods use the marginal maximum likelihood method and multiple EM cycles, but are different in that while the BILOG/Prior method calibrates the pretest items concurrently with the operational items, MEM calibrates only the pretest items.

Stocking's Method A

Stocking (1988) investigated two online calibration methods: Method A and Method B. Method A computes a maximum likelihood estimate of an examinee's ability using the item responses and parameter estimates for the operational items. The log-likelihood function of observing the responses, u_{ij}, on n operational items for N examinees with ability θ_j is

$$\ln L(U \mid \theta_1,\ldots,\theta_N) = \sum_{j=1}^{N}\sum_{i=1}^{n}\{u_{ij}\log P_i(\theta_j) + (1-u_{ij})\log[1-P_i(\theta_j)]\}. \tag{6}$$

Taking the derivative of Equation (6) with respect to the ability parameter yields

$$\begin{aligned}\frac{\partial}{\partial\theta_j}\ln L(U \mid \theta_1,\ldots,\theta_N) = &\\ &\sum_{i=1}^{n} u_{ij}\frac{1}{P_i(\theta_j)}\frac{\partial P_i(\theta_j)}{\partial\theta_j} + \\ &\sum_{i=1}^{n}(1-u_{ij})\frac{1}{1-P_i(\theta_j)}\frac{\partial(1-P_i(\theta_j))}{\partial\theta_j}.\end{aligned} \tag{7}$$

There would be one such derivative for each of the N examinees. For a given examinee, a maximization procedure (e.g., Newton-Raphson) can be performed to produce the maximum likelihood estimate of ability using the item responses on the operational items.

Stocking's Method A fixes the ability estimates (i.e., treats the ability estimates as the true abilities) obtained from Equation 7. The fixed abilities are then used to estimate item parameters for the pretest items. The log-likelihood function of observing the responses, x_{kj}, on a pretest item for N examinees given true abilities is

$$\ln L(X \mid \theta_1,\ldots,\theta_N) = \sum_{j=1}^{N} \log L(x_{kj} \mid \theta_j) = \sum_{j=1}^{N} x_{kj} \log P_k(\theta_j) + (1 - x_{kj}) \log[1 - P_k(\theta_j)]. \quad (8)$$

Taking the derivative of Equation 8 with respect to the item parameters for an item ($\Delta_k = a_k$, b_k, or c_k) yields three equations of the form

$$\frac{\partial}{\partial \Delta_k} \ln L(X \mid \theta_1,\ldots,\theta_N) = \sum_{j=1}^{N} x_{kj} \frac{1}{P_k(\theta_j)} \frac{\partial P_k(\theta_j)}{\partial \Delta_k} + \sum_{j=1}^{N} (1 - x_{kj}) \frac{1}{1 - P_k(\theta_j)} \frac{\partial (1 - P_k(\theta_j))}{\partial \Delta_k}. \quad (9)$$

For a given pretest item, a maximization procedure (e.g., Newton-Raphson) can be applied to produce the maximum likelihood estimate of item parameters using the item responses on the pretest items. Because the ability estimates are on the same scale as the operational item pool and the ability estimates are fixed in the calibration of the new items, rescaling of the item parameter estimates of the pretest items is not required.

The problem with Stocking's Method A is that it treats ability estimates obtained from the item responses on the operational items as true abilities in order to maintain the scales of subsequent item pools. Therefore, errors will be introduced in calibrating the pretest items because estimated abilities may be different from true abilities. Nevertheless, this method is a natural and simple way to calibrate the pretest items.

Stocking's Method B

Stocking's Method B (Stocking, 1988) is an enhanced version of Stocking's Method A. The method uses a set of previously calibrated 'anchor' items for scale maintenance to correct for scale drift that may result from the use of estimated abilities, rather than true abilities. The item parameter estimates of the anchor items are on the same scale as that of the operational items. Each examinee is administered some operational items, pretest items, and anchor items. As in Stocking's Method A, the ability estimate of each examinee is obtained using the operational item responses. The ability estimate is, then, fixed to calibrate the pretest items and anchor items. The two sets of item parameter estimates for the anchor items, the original item parameters and the re-estimated parameters, are used to compute a scale transformation to minimize the difference between the two test characteristic curves (Stocking & Lord, 1983). This scale transformation is then used to place the parameter estimates for the pretest items on the same scale as the item pool.

In using this method, it is important for the set of anchor items to be representative of the adaptive test item pool in terms of difficulty. Otherwise, inappropriate scale transformations derived from the anchor items would be applied to all the pretest items. The quality of the anchor items should be good, because poor anchor items could distort the scale transformation (Stocking, 1988, p. 20). The increase in the actual test length due to the inclusion of anchor items is a disadvantage of this method.

Method

Data

This study used nine 60-item ACT Mathematics test forms (ACT, 1997). Randomly equivalent groups of about 2,600 examinees took each form. The computer program BILOG (Mislevy & Bock, 1990) was used to estimate the item parameters for all items assuming a 3-PL logistic IRT model. These estimated item parameters are treated as population "true" item parameters, and they were used for generating simulated data. A total of 540 items were allocated as follows: 520 CAT operational item pool, 10 pretest items, and 10 anchor items. The 10 pretest items were randomly selected from the 540 items. The 10 anchor items were selected to be representative of the 520 operational items in terms of item difficulty.

CAT Simulation Procedure

Since true item parameters are never known in real world, this study used true item parameters only for generating item responses and for evaluating the performance of each item calibration method. We used estimated item parameters (hereafter referenced to as "baseline" parameter estimates) for item selection and ability estimation in CAT simulations. The baseline parameters for all 540 items were estimated from a full item-simulee response matrix generated using the true parameters and 3,000 randomly selected simulees from a standard normal distribution. The computer program BILOG-MG was used to concurrently calibrate the full response matrix.

Fixed-length adaptive tests (30 items) were administered to the randomly selected simulees with sample sizes of 300, 1,000, and 3,000 from the standard normal ability distribution. The 10 fixed-length "nonadaptive" pretest items were simulated using the same theta distribution. The data on pretest items here consisted of a full matrix, while the matrix of item responses on the operational CAT items was sparse.

The CATs were scored using Bock and Mislevy's (1982) expected a posteriori (EAP) ability estimation procedure. The initial prior distribution for EAP ability estimates was assumed to be normal with a mean of 0.0 and a standard deviation of 1.00. The simulated CAT began with an item of medium difficulty and used maximum information selection procedures thereafter.

At the end of the 30 fixed-length tests, ability estimates were computed using maximum likelihood estimation (MLE) procedures. This simulation was replicated 100 times for each method and sample size.

Pretest Item Calibration Procedures

The pretest items were calibrated and put on the same scale based on the methods described in the previous section. The computer simulations were done using programs written in Visual Basic and C++. An open-source C++ toolkit for IRT parameter estimation (Hanson, 2000) was used to implement the item parameter estimation for all methods except the BILOG/Prior method.

For the BILOG/Prior method, BILOG-MG (Zimowski, Muraki, Mislevy, & Bock, 1999) was used for pretest item calibrations[†]. Strong priors were based on the baseline parameter estimates described in the previous section. For each of the operational items in the pool, priors of each parameter were specified as follows: mean of log normal prior for a was the log of the baseline parameter estimate of a, standard deviation of log normal prior for a was 0.001; mean of normal prior for b was the baseline parameter estimate of b, standard deviation of normal prior for b was 0.005, alpha parameter for beta prior distribution for c was the value of ((1,000×the

[†] Since BILOG (Mislevy & Bock, 1990) and BILOG-MG (Zimowski et al., 1999) work in the same way for one-group, the results in this study would likely be the same if BILOG had been used instead of BILOG-MG.

baseline parameter estimate of c)+1), and beta parameter for beta prior distribution for c was the value of ((1,000× (1- the baseline parameter estimate of c))+1). This way of setting priors is recommended by the BILOG-MG manual for setting an item parameter to a fixed value (see Zimowski et al., 1999, pp. 89-90). The BILOG-MG default priors (c ~ Beta(5,17) and a ~ lognormal (0, 0.5)) were used for the pretest items.

BILOG-MG needs some responses on the operational items for calibration, even though strong priors were set on the operational CAT items. Due to the nature of the CAT administration, however, the operational items had a sparse item-simulee response matrix in which some items were never administered to simulees. In running BILOG-MG, we selected the operational CAT items that had at least 50 responses for the 300 sample size condition. Since many BILOG-MG runs with the minimum of 50 responses produced errors and stopped, we increased the minimum response (sample size) per item from 50 through 70 up to 100. We decided 100 responses per item to be a minimum for all different research conditions. Setting a high number for the minimum responses per item results in the small number of operational items available to compute the posterior distribution.

Using both the OEM and MEM methods, the operational items were completely fixed to be the same as the corresponding baseline parameter estimates. The same default priors on the pretest items as used for the BILOG/Prior method were set for comparison.

Stocking's Method A used the adaptively administered operational items to estimate the simulees' ability and treated them as true to calibrate pretest items. The same priors as the BILOG/Prior method were used for the pretest items.

When implementing Stocking's Method B, the item parameters obtained by Stocking's Method A were further transformed through the anchor items using the Stocking-Lord method (Stocking & Lord, 1983).

Criteria

In each condition (3 different sample sizes) studied, simulations are replicated 100 times. This produced 100 item parameter estimates for each of 10 pretest items in each condition for each method. The estimates were on the same scale as items in the operational item pool. The performance of the pretest item calibration methods was evaluated by the extent to which the true item characteristic curves of the pretest items were recovered.

Let $P(6 \mid a_k, b_k, c_k)$ be the true item characteristic curve for the 3-PL logistic item response model, where a_k, b_k, and c_k are the true item parameters for pretest item k. Let $P(\theta \mid \hat{a}_{kr}, \hat{b}_{kr}, \hat{c}_{kr})$ be the estimated item characteristic curve for item k on replication r, where $\hat{a}_{kr}, \hat{b}_{kr}, \hat{c}_{kr}$ are estimated pretest item parameters. An item characteristic curve criterion (Hanson & Béguin, 1999) for pretest item k is

$$\frac{1}{100}\sum_{r=1}^{100}\int_{-6}^{6}[P(\theta \mid a_k, b_k, c_k) - P(\theta \mid \hat{a}_{kr}, \hat{b}_{kr}, \hat{c}_{kr})]^2 w(\theta) d\theta, \qquad (10)$$

where $w(\theta)$ is a weight function based on a $N(0, 1)$ distribution. The integral is approximated using evenly spaced discrete θ points on the finite interval (-6, 6) at increments of 0.1. Each finite θ point was weighted by its probability under a normal distribution.

Equation 10 is the weighted mean squared difference between the true item characteristic curve and the estimated item characteristic curve, which is called the weighted mean squared error (WMSE). WMSE may be decomposed into the weighted squared bias (WSBias) and the weighted variance (WVariance):

$$\frac{1}{100}\sum_{r=1}^{100}\int_{-6}^{6}[P(\theta \mid a_k, b_k, c_k) - P(\theta \mid \hat{a}_{kr}, \hat{b}_{kr}, \hat{c}_{kr})]^2 w(\theta)d\theta =$$

$$\int_{-6}^{6}[P(\theta \mid a_k, b_k, c_k) - m_k(\theta)]^2 w(\theta)d\theta + \frac{1}{100}\sum_{r=1}^{100}\int_{-6}^{6}[P(\theta \mid \hat{a}_{kr}, \hat{b}_{kr}, \hat{c}_{kr}) - m_k(\theta)]^2 w(\theta)d\theta, \quad (11)$$

where

$$m_k(\theta) = \frac{1}{100}\sum_{r=1}^{100} P(\theta \mid \hat{a}_{kr}, \hat{b}_{kr}, \hat{c}_{kr}),$$

and WSBias and WVariance are the first and second terms on the right side of Equation 11. Also, average mean values and standard deviations of the WVariance, the WSBias, and the WMSE across pretest items in each condition were computed.

Results

The empirical results of the performance of the pretest item calibration/scaling methods appear in Tables 1 and 2 and Figures 1 through 3. WMSE, WSBias, and WVariance are presented in Table 1 for each pretest item, calibration method, and sample size. In addition, average values and standard deviations of the error indices over pretest items are presented in Table 1. Average WMSE along with the standard error of the average WMSE appear in Table 2. Figures 1 through 3 plot the means of error indices for different methods and sample sizes.

Many BILOG-MG runs were not successful under the 300 and 1,000 sample size conditions. Because of the sparseness of the data for the operational items, BILOG-MG very often produced errors in estimating item parameters and stopped, particularly in the 300 sample size condition. Under the 3,000 sample size condition, however, BILOG-MG worked properly. The results of BILOG-MG runs are provided in Tables 1 and 2 only for the 3,000 sample size condition.

Weighted Variance

The weighted variances of each of five methods are presented in Table 1 and the average weighted variances of each method are displayed in Figure 1. One expected result that holds for all methods was that the weighted variances decreased when the sample size increased. OEM method produced the smallest weighted variance under the 300 and 1,000 sample size conditions, Stocking's Method A the second smallest weighted variance, MEM method the third smallest weighted variance, and Stocking's Method B produced the largest weighted variance. For the 3,000 sample size condition, the BILOG/Prior method produced the largest weighted variance while the rank order of the weighted variances was the same for the other methods. However, differences in the weighted variance across methods appeared not to be substantial.

The results in Table 1 and Figure 1 show that although MEM utilized item response information on both operational and pretest items more intensively by taking multiple EM cycles than OEM did, it produced a larger weighted variance. The results also show that Stocking's Method B, which transforms the scale of pretest items that were calibrated by Stocking's Method A to the scale of operational items, resulted in a larger weighted variance than Stocking's Method A did.

Weighted Squared Bias

The weighted squared biases of each method for different sample size conditions are presented in Table 1, and the average weighted squared biases of each method for different sample size conditions are plotted in Figure 2. The weighted squared bias was less affected by the changes of sample size than the weighted variance across all methods. For example, for OEM, Stocking's Method A, and Stocking's Method B, there were only minor differences in the weighted biases between the 1,000 and 3,000 sample size conditions. In particular, Stocking's Method B produced similar weighted squared biases across all sample size conditions.

Comparing the results of the methods in terms of the weighted squared bias, the MEM method performed better than the other methods in this study. The MEM method produced the smallest weighted squared bias across all sample sizes. Even the weighted squared bias of the MEM method under the 300 sample size condition was smaller than the weighted squared biases of most methods for any sample size, except that of the BILOG/Prior method for a sample size of 3,000. The BILOG/Prior method worked well for the 3,000 sample size. The results of MEM and BILOG/Prior show that the methods using MMLE with multiple EM cycles tend to produce smaller weighted squared biases than other methods. Table 1 and Figure 2 indicate that Stocking's Method B produced the second smallest weighted squared biases across all sample sizes, except that of the BILOG/Prior method for the 3,000 sample size. The weighted squared biases of Stocking's Method B became slightly larger as sample sizes increased, which was not the case for the other methods. However, the absolute differences among the biases for the various sample sizes were small, so the differences may be due to sampling error. Stocking's Method A shows the largest weighted squared biases across all sample size conditions.

Unlike the weighted variance, MEM and Stocking's Method B produced obviously smaller weighted squared biases than OEM and Stocking's Method A, respectively, across all sample sizes.

Weighted Mean Squared Error

The weighted mean squared error of each method for different conditions are presented in Table 1. Table 2 presents the average weighted mean squared errors by the methods and sample sizes along with standard errors of the average weighted mean squared errors over replications.

Figure 3 plots a ± 1 standard error band around the means of the weighted mean squared errors, which is about a 68% confidence interval.

Figure 3 clearly shows that when BILOG/Prior is put aside, MEM tends to have the smallest WMSE, Stocking's Method B the second smallest WMSE, OEM the third smallest WMSE, and Stocking's Method A tends to produce the largest WMSE across all sample size conditions. The BILOG/Prior method under the 3,000 sample size performed similarly to MEM. It seems that since BILOG/Prior and MEM are mathematically the same (although they are different in actual implementation), the WMSEs of both methods were similar. The results in Figure 3 show that Stocking's Method B produced smaller total errors in parameter estimation than Stocking's Method A. This reflects that the scale transformation through anchor items is associated with decreases in total error for Stocking's Method B. The MEM method also produced smaller total error than OEM did.

The relative values of WMSEs for the methods were consistent across different sample size conditions. Some WMSE intervals for the methods within sample size conditions overlapped (see Figure 3). However, the MEM method always produced the smallest WMSE and Stocking's Method A produced the largest WMSE under the different sample size conditions studied.

Conclusion and Discussion

Our primary goal in this study was to compare the properties of five online pretest item calibration/scaling methods (MEM, OEM, Stocking's Method A, Stocking's Method B, and BILOG/Prior) under three different sample size conditions (300, 1,000, and 3,000). We expected that the results would provide CAT practitioners with guidance on which method(s) produce smaller parameter estimation error under small to large sample size conditions.

The MEM method produced the smallest total error (i.e., WMSE) in pretest item parameter estimation from small to large sample sizes. The BILOG/Prior method did not work appropriately under the 300 and 1,000 sample size conditions, but it performed similarly to MEM under the 3,000 sample size. Stocking's Method B produced the second smallest WMSE under the 300 and 1,000 sample size conditions (and the third smallest WMSE under the 3,000 sample size condition). The OEM method produced the next smallest WMSE. Stocking's Method A provided the largest total error.

The MEM method appears to be the best choice among the methods used in this study because it produced the smallest parameter estimation errors for all sample size conditions. Stocking's Method B also worked very well, but it requires anchor items that would make test lengths longer when other things are equal. The BILOG/Prior method did not seem to work with small sample sizes. Until more appropriate ways of handling sparse data with BILOG are devised, the BILOG/Prior method may not be a reasonable choice. Because Stocking's Method A has the largest weighted total error as well as a theoretical weakness (i.e., treating estimated ability as true ability), there appears to be little reason to use it. The MEM method should be preferred to OEM unless amount of time involved in iterative computation is a great concern. Otherwise, OEM and MEM are mathematically similar and OEM produces larger errors than the MEM method.

It is emphasized that the results reported here should be interpreted with caution due to the small to modest size of some of the reported error differences among the methods and the use of qualified pretest items. The pretest items used in this study were actually operational items, so the quality of the items was relatively high. When the pretest items are poor or do not fit the 3PL model, the performance of the methods considered in this study may be different. In practice, OEM may perform better than MEM when there are some bad pretest items. Further research should look at performance of methods when some pretest items are bad and there is a sparse matrix of item responses for pretest items.

References

ACT, Inc. (1997). *ACT assessment technical manual.* Iowa City, IA: Author.

Bock, R. D., & Mislevy, R. J. (1982). Adaptive EAP estimation of ability in microcomputer environment. *Applied Psychological Measurement, 6*, 431-444.

Folk, V.G., & Golub-Smith, M. (1996). *Calibration of on-line pretest data using BILOG.* Paper presented at the annual meeting of the National Council of Measurement in Education, New York.

Hanson, B. A. (2000). *Estimation toolkit for item response models (ETIRM).* (Available at http://www.b-a-h.com/software/cpp/etirm.html).

Hanson, B. A., & Béguin, A. A. (1999). *Separate versus current estimation of IRT item parameters in the common item equating design* (Research Report 99-8). Iowa City, IA: ACT, Inc.

Haynie, K. A., & Way, W. D. (1995). *An investigation of item calibration procedures for a computerized licensure examination.* Paper presented at symposium entitled Computer Adaptive Testing, at the annual meeting of NCME, San Francisco.

Hsu, Y., Thompson, T. D., & Chen, W.-H. (1998). *CAT item calibration.* Paper presented at the Annual Meeting of the National Council on Measurement in Education, San Diego.

Levine, M. V., & Willams, B. A. (1998). *Development and evaluation of online calibration procedures* (TCN #96-216). Champaign, IL: Algorithm Design and Measurement Services, Inc.

Mislevy, R. J., & Bock, R. J. (1990). *BILOG3: Item analysis and test scoring with binary logistic model* (2nd ed.) [Computer program]. Mooresville, IN: Scientific Software.

Parshall, C. G. (1998). *Item development and pretesting in a computer-based testing environment.* Paper presented at the colloquium, Computer-Based Testing: Building the Foundation for Future Assessments, Philadelphia, PA.

Samejima, F. (2000). *Some considerations for improving accuracy of estimation of item characteristic curves in online item calibration of computerized adaptive testing.* Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans.

Spray, J. A., Parshall, C. G., & Huang, C.-H. (1997). *Calibration of CAT item administered online for classification: Assumption of local independence.* Paper presented at the annual meeting of the Psychometric Society, Gatlinberg, TN.

Stocking, M. L. (1988). *Scale drift in on-line calibration* (Research Report 88-28). Princeton, NJ: ETS.

Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement, 7*, 201-210.

Wainer, H., & Mislevy, R. J. (1990). Item response theory, item calibration, and proficiency estimation. In Wainer, H. (Ed.), *Computer adaptive testing: A primer* (Chapter 4, pp. 65-102). Hillsdale, NJ: Lawrence Erlbaum.

Zimowski, M. F., Muraki, E., Mislevy, R. J., & Bock, R. D. (1999). *BILOG-MG: Multiple-group IRT analysis and test maintenance for binary items* [Computer program]. Chicago: Scientific Software International.

TABLE 1

Weighted Mean Squared Error, Weighted Squared Bias, and Weighted Variance for Pretest Item Calibration Methods

Calibration			Sample300			Sample1000			Sample3000	
Methods	Pretest Items	WMSE	WSBias	WVariance	WMSE	WSBias	WVariance	WMSE	WSBias	WVariance
	Item1	1.61E-03	5.35E-04	1.07E-03	7.56E-04	3.93E-04	3.63E-04	5.02E-04	3.57E-04	1.45E-04
	Item2	1.65E-03	4.84E-04	1.17E-03	6.78E-04	3.20E-04	3.57E-04	4.65E-04	3.41E-04	1.24E-04
	Item3	1.12E-03	4.14E-04	7.04E-04	8.12E-04	5.29E-04	2.83E-04	5.61E-04	4.49E-04	1.12E-04
	Item4	1.65E-03	3.18E-04	1.33E-03	6.13E-04	2.01E-04	4.12E-04	3.88E-04	2.52E-04	1.36E-04
	Item5	1.36E-03	2.00E-04	1.16E-03	5.14E-04	8.40E-05	4.30E-04	3.11E-04	1.40E-04	1.71E-04
OEM	Item6	1.22E-03	3.64E-05	1.19E-03	4.39E-04	7.46E-05	3.64E-04	2.65E-04	9.22E-05	1.73E-04
	Item7	1.45E-03	3.15E-04	1.14E-03	6.52E-04	2.50E-04	4.02E-04	3.94E-04	2.52E-04	1.42E-04
	Item8	1.23E-03	2.11E-05	1.21E-03	4.82E-04	6.91E-05	4.13E-04	2.34E-04	9.45E-05	1.39E-04
	Item9	8.39E-04	1.62E-04	6.78E-04	3.87E-04	1.72E-04	2.15E-04	2.38E-04	1.74E-04	6.42E-05
	Item10	1.49E-03	4.28E-04	1.06E-03	6.62E-04	2.47E-04	4.16E-04	3.52E-04	1.97E-04	1.55E-04
	Mean	1.36E-03	2.92E-04	1.07E-03	6.00E-04	2.34E-04	3.66E-04	3.71E-04	2.35E-04	1.36E-04
	SD*	2.62E-04	1.81E-04	2.14E-04	1.39E-04	1.49E-04	6.83E-05	1.13E-04	1.19E-04	3.16E-05
	Item1	1.43E-03	2.88E-04	1.14E-03	5.29E-04	1.35E-04	3.94E-04	2.14E-04	5.24E-05	1.62E-04
	Item2	1.36E-03	7.13E-05	1.28E-03	4.11E-04	4.12E-06	4.07E-04	1.46E-04	3.93E-06	1.42E-04
	Item3	8.09E-04	4.24E-05	7.66E-04	3.57E-04	3.75E-05	3.20E-04	1.30E-04	3.28E-06	1.27E-04
	Item4	1.49E-03	4.70E-05	1.45E-03	5.05E-04	1.45E-05	4.90E-04	1.58E-04	2.08E-06	1.56E-04
	Item5	1.30E-03	3.17E-05	1.27E-03	5.07E-04	1.50E-05	4.92E-04	2.04E-04	6.42E-06	1.98E-04
MEM	Item6	1.29E-03	3.15E-05	1.26E-03	4.10E-04	1.59E-05	3.94E-04	1.92E-04	5.99E-06	1.86E-04
	Item7	1.27E-03	3.41E-05	1.23E-03	4.60E-04	3.78E-06	4.56E-04	1.67E-04	2.73E-06	1.64E-04
	Item8	1.30E-03	1.67E-05	1.28E-03	4.33E-04	2.10E-06	4.31E-04	1.49E-04	7.27E-08	1.49E-04
	Item9	7.97E-04	1.89E-05	7.79E-04	2.59E-04	7.74E-06	2.51E-04	7.59E-05	2.91E-06	7.30E-05
	Item10	1.27E-03	1.11E-04	1.16E-03	4.95E-04	1.29E-05	4.82E-04	1.86E-04	3.25E-06	1.83E-04
	Mean	1.23E-03	6.92E-05	1.16E-03	4.37E-04	2.49E-05	4.12E-04	1.62E-04	8.31E-06	1.54E-04
	SD	2.37E-04	8.18E-05	2.22E-04	8.26E-05	4.00E-05	7.81E-05	4.07E-05	1.56E-05	3.56E-05

TABLE 1 *(continued)*

Calibration Methods	Pretest Items	Sample300 WMSE	Sample300 WSBias	Sample300 WVariance	Sample1000 WMSE	Sample1000 WSBias	Sample1000 WVariance	Sample3000 WMSE	Sample3000 WSBias	Sample3000 WVariance
Stocking's Method A	Item1	1.71E-03	5.57E-04	1.15E-03	8.37E-04	4.42E-04	3.94E-04	6.09E-04	4.51E-04	1.58E-04
	Item2	1.80E-03	5.15E-04	1.28E-03	7.67E-04	3.68E-04	3.99E-04	5.09E-04	3.75E-04	1.34E-04
	Item3	1.46E-03	6.76E-04	7.83E-04	1.13E-03	8.17E-04	3.11E-04	7.95E-04	6.80E-04	1.16E-04
	Item4	1.75E-03	3.05E-04	1.44E-03	6.51E-04	1.95E-04	4.56E-04	3.98E-04	2.50E-04	1.48E-04
	Item5	1.49E-03	2.37E-04	1.26E-03	5.76E-04	1.11E-04	4.65E-04	3.62E-04	1.78E-04	1.83E-04
	Item6	1.31E-03	7.14E-05	1.24E-03	5.22E-04	1.32E-04	3.90E-04	3.44E-04	1.62E-04	1.82E-04
	Item7	1.55E-03	3.09E-04	1.24E-03	7.00E-04	2.59E-04	4.41E-04	4.20E-04	2.65E-04	1.55E-04
	Item8	1.35E-03	6.04E-05	1.29E-03	5.97E-04	1.65E-04	4.32E-04	3.43E-04	1.97E-04	1.46E-04
	Item9	1.15E-03	5.01E-04	6.47E-04	8.08E-04	5.92E-04	2.16E-04	6.54E-04	5.87E-04	6.67E-05
	Item10	1.58E-03	4.40E-04	1.14E-03	7.69E-04	3.18E-04	4.51E-04	3.96E-04	2.29E-04	1.67E-04
	Mean	1.51E-03	3.67E-04	1.15E-03	7.36E-04	3.40E-04	3.96E-04	4.83E-04	3.37E-04	1.45E-04
	SD	2.06E-04	2.06E-04	2.45E-04	1.73E-04	2.25E-04	7.78E-05	1.55E-04	1.81E-04	3.45E-05
Stocking's Method B	Item1	1.52E-03	3.22E-04	1.20E-03	6.55E-04	2.27E-04	4.28E-04	3.81E-04	2.17E-04	1.65E-04
	Item2	1.49E-03	1.55E-04	1.33E-03	5.93E-04	1.32E-04	4.61E-04	2.92E-04	1.43E-04	1.49E-04
	Item3	1.07E-03	1.66E-04	9.06E-04	7.12E-04	3.47E-04	3.66E-04	4.48E-04	3.06E-04	1.42E-04
	Item4	1.51E-03	4.26E-05	1.47E-03	5.65E-04	9.26E-05	4.73E-04	2.33E-04	8.10E-05	1.52E-04
	Item5	1.38E-03	5.31E-05	1.33E-03	5.21E-04	3.32E-05	4.87E-04	2.37E-04	4.41E-05	1.93E-04
	Item6	1.38E-03	6.26E-05	1.32E-03	4.43E-04	2.66E-05	4.16E-04	2.38E-04	4.83E-05	1.90E-04
	Item7	1.31E-03	9.71E-05	1.21E-03	5.77E-04	8.99E-05	4.87E-04	2.20E-04	6.44E-05	1.56E-04
	Item8	1.32E-03	1.41E-05	1.30E-03	5.16E-04	5.40E-05	4.62E-04	2.26E-04	6.93E-05	1.57E-04
	Item9	9.02E-04	1.11E-04	7.90E-04	4.81E-04	2.26E-04	2.55E-04	3.61E-04	2.81E-04	7.96E-05
	Item10	1.40E-03	1.93E-04	1.21E-03	5.71E-04	7.03E-05	5.00E-04	2.71E-04	9.06E-05	1.81E-04
	Mean	1.33E-03	1.22E-04	1.21E-03	5.63E-04	1.30E-04	4.33E-04	2.91E-04	1.34E-04	1.56E-04
	SD	1.98E-04	9.12E-05	2.07E-04	7.98E-05	1.04E-04	7.47E-05	7.91E-05	9.85E-05	3.21E-05

TABLE 1 *(continued)*

Calibration Methods	Pretest Items	Sample300 WMSE	Sample300 WSBias	Sample300 WVariance	Sample1000 WMSE	Sample1000 WSBias	Sample1000 WVariance	Sample3000 WMSE	Sample3000 WSBias	Sample3000 WVariance
	Item1							2.67E-04	8.03E-05	1.86E-04
	Item2							1.91E-04	1.93E-05	1.71E-04
	Item3							1.66E-04	8.11E-06	1.58E-04
	Item4							1.96E-04	7.27E-06	1.88E-04
	Item5							2.26E-04	7.73E-06	2.19E-04
BILOG/Prior	Item6	NA	NA	NA	NA	NA	NA	2.47E-04	3.93E-05	2.08E-04
	Item7							2.05E-04	2.12E-05	1.84E-04
	Item8							1.78E-04	1.16E-05	1.67E-04
	Item9							8.42E-05	3.26E-06	8.10E-05
	Item10							2.16E-04	9.55E-06	2.06E-04
	Mean							1.98E-04	2.08E-05	1.77E-04
	SD							5.02E-05	2.34E-05	3.87E-05

* Standard deviation over pretest items; NA = Not available because BILOG-MG stopped inappropriately
WMSE = Weighted Mean Squared Error; WSBias = Weighted Squared Bias; WVariance = Weighted Variance

TABLE 2

Mean of Weighted Mean Squared Error

Calibration Methods	Sample Size 300	Sample Size 1000	Sample Size 3000
OEM	1.36E-03 (4.07E-04)	6.00E-04 (1.59E-04)	3.71E-04 (8.13E-05)
MEM	1.23E-03 (3.90E-04)	4.37E-04 (1.14E-04)	1.62E-04 (4.65E-05)
Stocking's Method A	1.51E-03 (4.53E-04)	7.36E-04 (1.89E-04)	4.83E-04 (9.99E-05)
Stocking's Method B	1.33E-03 (4.06E-04)	5.63E-04 (1.72E-04)	2.91E-04 (8.55E-05)
BILOG/Prior	NA	NA	1.98E-04 (6.16E-05)

( ) Standard error over replications

FIGURE 1. **Average Weighted Variance for Each Method**

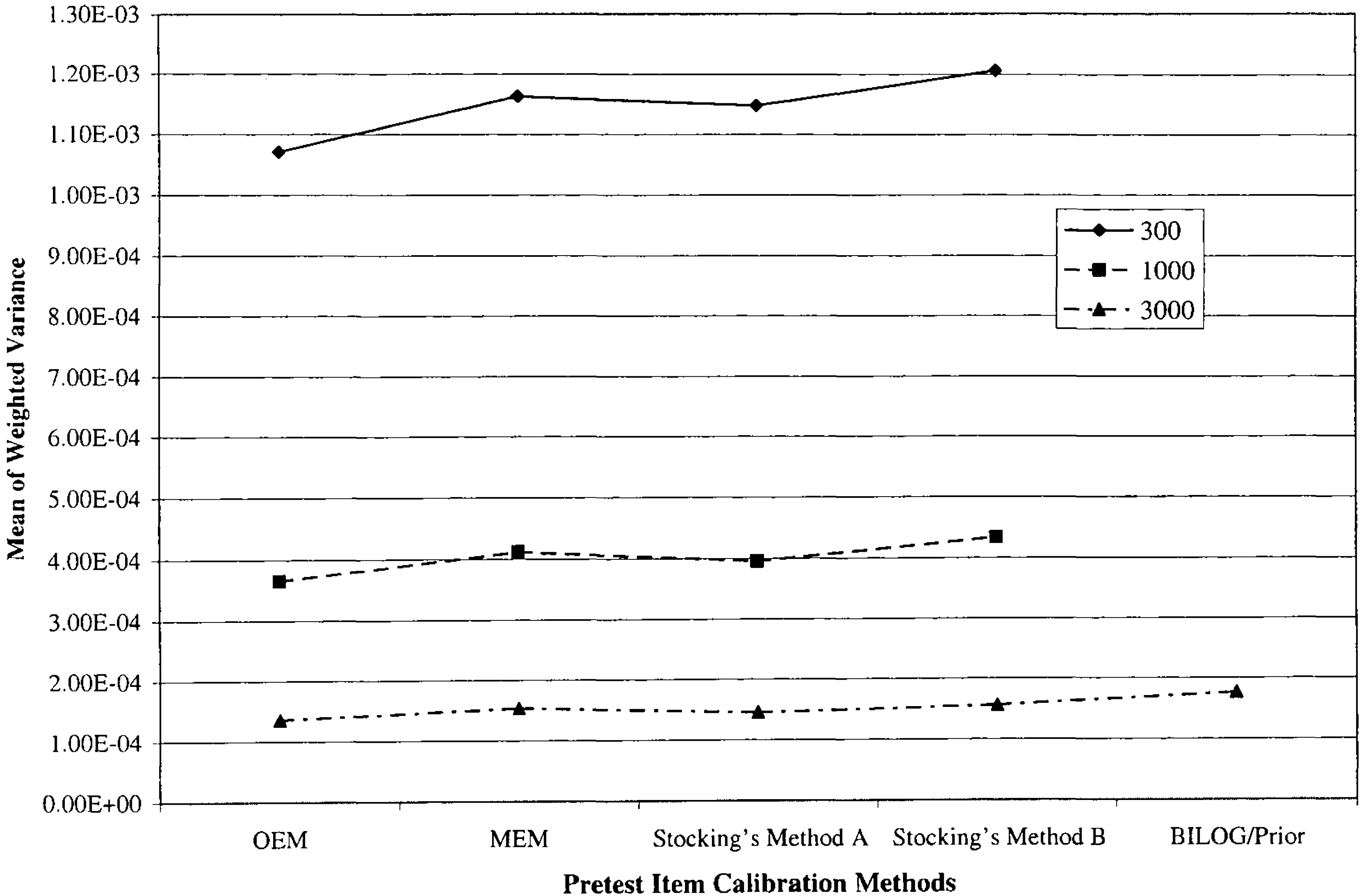

FIGURE 2. **Average Weighted Squared Bias for Each Method**

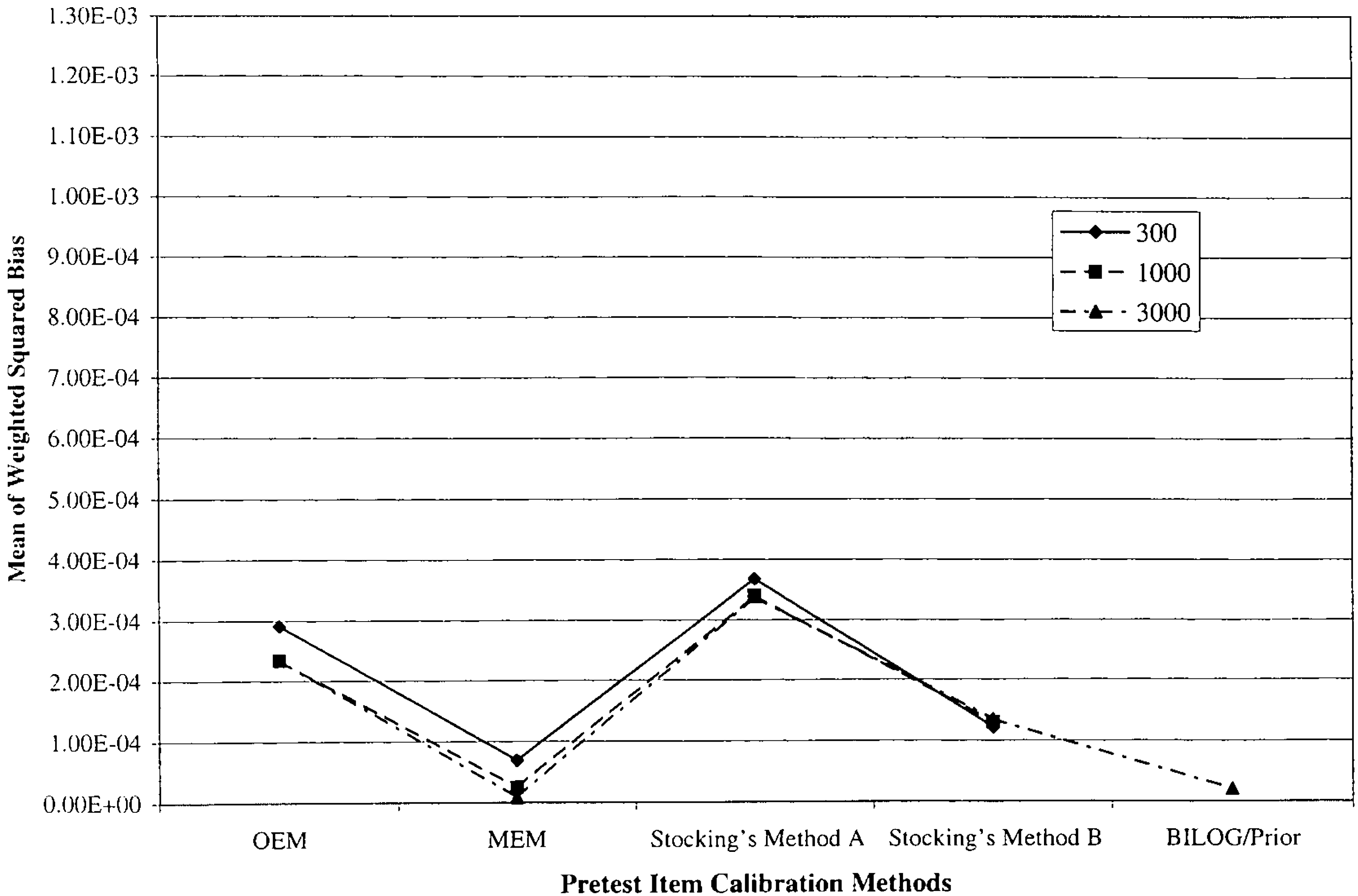

FIGURE 3. About a 68% Confidence Interval Around the Average Weighted Mean Squared Errors for Each Method