ACT Research Report Series 99-8

Separate Versus Concurrent Estimation of IRT Item Parameters in the Common Item Equating Design

Bradley A. Hanson

Anton A. Béguin

ACT

December 1999

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

Separate Versus Concurrent Estimation of IRT Item Parameters in the Common Item Equating Design

Bradley A. Hanson
ACT, Inc.

Anton A. Béguin
University of Twente

Abstract

IRT item parameters can be estimated using data from a common item equating design either separately for each form, or concurrently across forms. This paper reports the results of a simulation study of separate versus concurrent item parameter estimation. Using simulated data from a test with 60 dichotomous items, four factors were considered: 1) program (MULTILOG versus BILOG-MG), 2) sample size per form (3000 versus 1000), 3) number of common items (20 versus 10), and 4) equivalent versus nonequivalent groups taking the two forms (no mean difference versus a mean difference of 1 standard deviation). In addition, four methods of item parameter scaling were used in the separate estimation condition: two item characteristic curve methods (Stocking-Lord and Haebara), and two moment methods (Mean/Mean and Mean/Sigma). Although concurrent estimation resulted in less error than separate estimation more times than not, it is argued that the results of this study, together with other research on this topic, are not sufficient to recommend completely avoiding separate estimation in favor of concurrent estimation.

Separate Versus Concurrent Estimation of IRT Item Parameters in the Common Item Equating Design

The latent variable in many IRT (item response theory) models is unidentified up to a linear transformation. This means that if the latent variable is linearly transformed then an appropriate linear transformation can be made to the item parameters so that the model produces exactly the same fitted probabilities. In practice, a scale or metric for the IRT latent variable and estimated item parameters is determined by constraints imposed by the software used for parameter estimation. For the software studied in this paper the scale of the latent variable is determined by assuming the mean and standard deviation of the latent variable distribution used in the marginal maximum likelihood estimation are 0 and 1, respectively.

For example, the probability of a correct response for a dichotomous item i given by the three-parameter logistic IRT model (Lord, 1980) at latent variable value θ is

$$P(\theta \mid a_i, b_i, c_i) = c_i + \frac{1 - c_i}{1 + e^{-1.7a_i(\theta - b_i)}}, \tag{1}$$

where a_i, b_i, and c_i are item parameters for item i. Let the latent variable be linearly transformed by $\theta^* = A\theta + B$, let $a_i^* = a_i/A$, and let $b_i^* = Ab_i + B$. Substituting a_i^* for a_i, b_i^* for b_i, and θ^* for θ in Equation 1 will produce exactly the same probability of a correct response to item i as using a_i, b_i, and θ_i. For any linear transformation of the latent variable, a corresponding linear transformation of the item parameters a_i and b_i for any item i can be found to produce exactly the same probability of a correct response. Thus, the scale and location of the latent variable in the three-parameter logistic model are unidentified (any linear transformation of the latent variable produces exactly the same model fit).

In the common item nonequivalent groups equating design two forms of a test with some items in common are administered to samples from two populations. This paper compares two alternative procedures for producing item parameter estimates on a common scale in a common item nonequivalent groups equating design: concurrent and separate estimation. In concurrent estimation item parameters for all items on both forms are estimated simultaneously in one run of the estimation software. Estimating parameters for all items simultaneously assures that all parameter estimates are on the same scale. Estimation software that can handle multiple groups of examinees is required to properly perform concurrent estimation in the nonequivalent groups design.

In separate estimation the item parameters for the two forms are estimated using two separate runs of the estimation software. The item parameter estimates for the two forms will not be on the same scale. This is due to the fact that constraining the scale of the latent variable by fixing the mean and standard deviation of the latent variable distribution will result in different scales when samples from different populations are used for item parameter estimation. In separate estimation the two sets of item parameter estimates for the common items are used to estimate a scale transformation that will put the item parameter estimates of one form on the scale of the item parameter estimates for the other form. Several IRT scale transformation methods are available (Kolen and Brennan, 1995, Chapter 6).

Item parameter scaling is not needed when the groups taking the two forms are samples from the same population (equivalent groups). This study also includes a condition in which equivalent groups take the two forms. In the equivalent groups case the concurrent and separate estimation procedures are distinguished only by whether the item parameter estimation involves one or two runs of the estimation software. Even though item parameter scaling is not needed when equivalent groups are used it can still be performed. Item parameter scaling with equivalent groups may reduce estimation error by adjusting for small differences in the latent variable scale between the samples taking the two forms that are due to sampling error.

Little research has been done comparing the concurrent and separate estimation procedures. Petersen, Cook, and Stocking (1983) and Wingersky, Cook, and Eignor (1987) both concluded that concurrent estimation performed somewhat better than separate estimation. Both these studies used the computer program LOGIST (Wingersky, Barton, and Lord, 1982) which uses joint maximum likelihood to estimate the item parameters.

Kim and Cohen (1998) studied separate versus concurrent estimation with simulated data using the computer programs BILOG (Mislevy and Bock, 1990) and MULTILOG (Thissen, 1991) for item parameter estimation. Kim and Cohen (1998) concluded that separate and concurrent estimation provided similar results except when the number of common items was small, in which case separate estimation provided more accurate results. One limitation of their study is that BILOG was used for separate estimation and MULTILOG was used for concurrent estimation. Thus, differences between separate and concurrent estimation in the case of nonequivalent groups was confounded with the difference between computer programs. BILOG was also used for concurrent estimation, although this is not strictly appropriate in the case in which the groups taking the two forms are not randomly equivalent because BILOG cannot estimate the correctly specified model in which separate latent variable distributions are assumed for the groups of examinees taking the two forms.

Previous research has come to differing conclusions concerning the relative performance of separate and concurrent estimation. The objective of this paper is to provide further information concerning the relative performance of concurrent versus separate estimation for some conditions that have not been previously studied. In this paper BILOG-MG (Zimowski, Muraki, Mislevy, and Bock, 1996) and MULTILOG are used for both concurrent and separate estimation, so that unlike Kim and Cohen (1998) the difference between the results for concurrent and separate estimation will not be confounded with computer program when nonequivalent groups take the two forms. In addition, two distinct forms with common items are used, and multiple methods of item parameter scaling are examined (Kim and Cohen, 1998, used only one form and examined only one method of item parameter scaling).

Data

This study uses items from two 60 item ACT Assessment (ACT, 1997) Mathematics forms denoted forms A and Z. Randomly equivalent groups of 2696 and 2670 examinees took forms A and Z, respectively. The computer program BILOG (Mislevy & Bock, 1990) was used with these data to estimate the item parameters for all items assuming a three parameter logistic IRT model (Equation 1). These estimated item parameters were treated as population item parameters for simulating data.

The two forms do not have any items in common. Items on each of these two forms were divided into three sets of 20 items, such that the content and statistical characteristics of the three sets were as similar as possible. Form B was created by using the first set of 20 items from form A and the last two sets of 20 items from form Z. Thus, forms A and B have 20 items in common. The 20 items in common are considered an internal anchor. In this paper form A is considered the base form and form B is the new form — the form B item parameter estimates will be put on the scale of the form A item parameter estimates. The population item parameters for forms A and B are presented in Table 1. Form A consists of items 1 through 60, and form B consists of items 41 through 100, where items 41 through 60 are common to forms A and B.

Method

Samples of item responses for form A were generated by sampling the IRT latent variable (θ) from a normal distribution with mean zero and standard deviation one — denoted N(0,1). Two sets of item responses were generated for form B by sampling θ from a N(0,1) distribution and a N(1,1) distribution. The samples of form A and the mean 0 form B samples were used to examine the case of a common item equivalent groups design in which the samples administered the two forms are from the same population. The samples of form A and the mean 1 form B samples were used to examine the case of a common item nonequivalent groups design in which the samples administered the two forms are from different populations.

Fifty form A samples were generated, and 50 of each of the two sets of form B samples were generated for a total of 150 samples. For the 50 pairs of form A and mean 0 form B samples, and for the 50 pairs of form A and mean 1 form B samples, item parameters were estimated both separately for each form and simultaneously for both forms together. Two programs were used for item parameter estimation: BILOG-MG and MULTILOG. Both BILOG-MG and MULTILOG can handle the case where nonequivalent groups take forms A and B. Appendices A and B give the BILOG-MG and MULTILOG control files used to obtain parameter estimates for each simulated sample. Priors on the a and c parameters were used in MULTILOG to make the MULTILOG results more comparable with the BILOG-MG results for which priors were used on the a and c parameters by default. A difference between the models fit by BILOG-MG and MULTILOG for concurrent estimation in the form B mean 1 conditions was that the standard deviations of the latent distributions for the two groups were allowed to differ in BILOG-MG, but were not allowed to differ in MULTILOG. For both programs the means of the latent variable for the two groups were allowed to differ.

In the separate estimation conditions the scale of the item parameters for form B were put on the scale of the item parameters for form A using four item parameter scaling methods described in Kolen and Brennan (1995): 1) Mean/Mean, 2) Mean/Sigma, 3) Stocking-Lord, and 4) Haebara. The Mean/Mean and Mean/Sigma methods use moments of the item parameter estimates to produce a scale transformation (these will be referred to as moment methods), and the Stocking-Lord and Haebara methods minimize differences between item characteristic or test characteristic curves to produce a scale transformation (these will be referred to as characteristic curve methods). Item parameter scaling was performed when the groups taking the two forms were equivalent or nonequivalent, even though item parameter scaling is strictly only needed when the groups taking the two forms are nonequivalent. In the separate estimation conditions there will be two sets of item parameter estimates for the common items. In this study the form A (base form) item parameter estimates were used as the parameter estimates of the common items for the purpose of computing the criteria used to evaluate the quality of the item parameter scaling. This is in contrast to Kim and Cohen (1998), where the average of the item parameter estimates for the two forms were used as parameter estimates for the common items for the purpose of computing criteria.

Two levels of sample size were considered: 1) 3000 examinees per form, and 2) 1000 examinees per form. The 1000 sample size condition used the first 1000 of the 3000 examinees per sample. Two levels of the number of common items were used: 1) 20 items, and 2) 10 items. Ten of the common items in the 20 item condition were also considered common in the 10 item condition (items 42, 44, 45, 47, 48, 50, 51, 53, 55, and 59 in Table 1). The other 10 original common items were treated as unique items on the two forms in the 10 item condition. These 10 items were treated as separate sets of 10 items on forms A and B. The split of the original 20 common items into two 10 item sets was done such that the statistical characteristics of the two sets of 10 items were as similar as possible.

Five factors are investigated in this study: 1) equivalent versus nonequivalent groups administered the two forms, 2) concurrent versus separate estimation using four item parameter scaling methods (this factor has five levels), 3) estimation program (BILOG-MG and MULTILOG), 4) sample size (3000 and 1000), and 5) number of common items (20 and 10). There are a total of 80 conditions studied ($2 \times 5 \times 2 \times 2 \times 2$). For each of these conditions 50 sets of item parameter estimates for the items on the two forms were computed using the 50 samples.

Criteria

In each condition studied there are 50 sets of item parameter estimates for form A and a corresponding 50 sets of item parameter estimates for form B. In each of the 50 replications for each condition the form B item parameter estimates should be on the same scale as the population item parameters, after a transformation of the form B estimates in the separate estimation conditions. The extent to which this holds is assessed by two criteria: 1) a criterion based on the true score equating function from form B to form A;

and 2) a criterion based on how close the estimated item characteristic curves are to the true item characteristic curves for the form B items.

The first criterion is based on the true score equating function from true number correct scores on form B to true number correct scores on form A (Kolen and Brennan, 1995). Let τ_i be the form A true number correct score corresponding to an integer number correct score of i on form B. The mapping from i to τ_i uses the true score equating function from form B to form A. The true score equating function depends on the test characteristic curves for forms A and B, which in turn depend on the population item parameters for forms A and B. The criterion only evaluates the true score equating function from form B to form A at integer number correct scores on form B. These are the points used when applying the true score equating function to equate observed number correct scores on form B to number correct scores on form A. Note that τ_i is only defined for i greater than the sum of the three parameter logistic model c parameters for items on form B, and for i less than the number of items on form B.

Let t_{ij} be the estimated true score equating function corresponding to score i using item parameter estimates for forms A and B from replication j. The criterion based on the true score equating function at score i is

$$\frac{1}{50}\sum_{j=1}^{50}(\tau_i - t_{ij})^2 \,. \tag{2}$$

Equation 2 is a Monte Carlo estimate of the mean (expected) squared error of the true score equating function at score i with the expectation taken over the random variable representing the estimated true score equating function at score i. The estimated true score equating function depends on the random variables representing the item responses on forms A and B. Equation 2 can be written as

$$\frac{1}{50}\sum_{j=1}^{50}(\tau_i - t_{ij})^2 = (\tau_i - \bar{t}_i)^2 + \frac{1}{50}\sum_{j=1}^{50}(t_{ij} - \bar{t}_i)^2 \,, \tag{3}$$

where

$$\bar{t}_i = \sum_{j=1}^{50} t_{ij} \,.$$

In Equation 3 the mean squared error is written as a sum of the squared bias (the first term on the right hand side of Equation 3) and variance (the second term on the right hand side of Equation 3). The mean squared error represents the total error in the estimated true score equating function at score i. The squared bias and variance represent the portions of the mean squared error corresponding to systematic and random error.

The mean squared error, squared bias, and variance as given in Equation 3 are computed for each condition studied at 49 form B number correct scores $i = 11, 12, \ldots, 59$. These are the number correct scores on form B for which the true score equating function can be computed based on the population item parameters. There could be some replications where the true score equating function cannot be computed using estimated item parameters for certain form B number correct scores, even though the true score equating function can be computed using the true item parameters. Consequently, for some number correct scores i the sums in Equation 3 could involve less than 50 terms since only replications where the true score equating function is defined at score i can be used in the sums in Equation 3. It turned out that the true score equating function could be computed for all score points in each of the 50 replications in all conditions.

Two averages of the mean squared error, squared bias, and variance across form B number correct scores of 11 through 59 are computed for each condition. One average is an unweighted average across the scores. The second average is a weighted average with the weight given to each score being the population probability of that score on form B computed using the population item parameters and a $N(0, 1)$ latent variable distribution. Since only 49 of the 61 number correct scores are used in the average the probabilities are standardized so they sum to one over the number correct scores 11 through 59. Thus, a weighted and unweighted average of mean squared error, squared bias, and variance are computed for each condition.

The second criterion used to evaluate the results is a measure of the difference between the estimated and true item characteristic curves for the 60 form B items. Only the form B items are used in this criterion since the focus of this paper is comparing how well the new form (form B) item parameter estimates are properly put on the base (form A) scale. The item characteristic curve criterion for item i is

$$\frac{1}{50}\sum_{j=1}^{50}\int_{-6}^{6}[P(\theta \mid a_i, b_i, c_i) - P(\theta \mid \hat{a}_{ij}, \hat{b}_{ij}, \hat{c}_{ij})]^2 w(\theta) d\theta\,, \qquad (4)$$

where $P(\theta \mid a, b, c)$ is the item characteristic curve for the three parameter logistic item response model (Equation 1), a_i, b_i, c_i are the population item parameters for item i, $\hat{a}_{ij}, \hat{b}_{ij}, \hat{c}_{ij}$ are estimated item parameters for item i from replication j, and $w(\theta)$ is a weight function. The integral is taken over a finite interval $(-6, 6)$ in order to assure the integral is finite for all weight functions. When using the three parameter logistic model the integral over the whole real line may be infinite for some weight functions. Equation 4 is a Monte Carlo estimate of the mean (expected) squared error of the estimated item characteristic function with the expectation taken over the random variables representing the item parameter estimates, which depend on the random variables representing the item responses. Equation 4 can be written as

$$\frac{1}{50}\sum_{j=1}^{50}\int_{-6}^{6}[P(\theta \mid a_i, b_i, c_i) - P(\theta \mid \hat{a}_{ij}, \hat{b}_{ij}, \hat{c}_{ij})]^2 w(\theta)d\theta =$$

$$\int_{-6}^{6}[P(\theta \mid a_i, b_i, c_i) - m_i(\theta)]^2 w(\theta)d\theta + \frac{1}{50}\sum_{j=1}^{50}\int_{-6}^{6}[P(\theta \mid \hat{a}_{ij}, \hat{b}_{ij}, \hat{c}_{ij}) - m_i(\theta)]^2 w(\theta)d\theta, \quad (5)$$

where

$$m_i(\theta) = \frac{1}{50}\sum_{j=1}^{50} P(\theta \mid \hat{a}_{ij}, \hat{b}_{ij}, \hat{c}_{ij}).$$

The first term on the right side of Equation 5 is the squared bias of the estimated item characteristic curve and the second term on the right hand side of Equation 5 is the variance of the estimated item characteristic curve. Two values of mean squared error, squared bias, and variance are computed for each of the 60 form B items in each condition using two weight functions. The first weight function is $w(\theta) = 1$ (equal weighting). The second weight function is the density function for a $N(0,1)$ distribution. Since analytical expressions for the integrals in Equation 5 are not available Monte Carlo integration is used to compute these integrals by drawing 1000 uniformly distributed random variables from the interval -6 to 6. For the normal weight function each random deviate was weighted by its probability under a normal distribution. Average values of the mean squared error, squared bias, and variance of the item characteristic curves over the 60 items are computed for each condition. This will provide a summary of the total error, squared bias, and variance of the item parameter estimates for form B in each condition.

Results

Using the command files given in Appendix A and Appendix B most of the MULTILOG and BILOG-MG runs converged. Despite increasing the number of EM cycles to 100, most of the MULTILOG runs in the form B mean 1 concurrent estimation conditions did not converge. For the MULTILOG runs that did not converge, the final convergence criterion was close to the 0.001 cutoff for convergence. In one of the 50 samples in the 1000 sample size concurrent estimation conditions for BILOG-MG the EM steps converged, but divergence occurred during the Newton steps. Despite the divergence in the Newton steps, all item parameter estimates where within reasonable bounds.

Values of the weighted and unweighted average true score equating criterion, and the weighted and unweighted average ICC criterion are presented in Figures 1 through 4, respectively. The values plotted for the ICC criterion are the values from Equation 5 multiplied by 1000. In each figure there are 12 plots arranged in three rows and four columns. The plots in the three rows give results for the squared bias, variance, and mean squared error (MSE). The plots in the four columns give results for: 1) a sample size of 1000 with 10 common items, 2) a sample size of 1000 with 20 common items, 3) a sample size of 3000 with 10 common items, and 4) a sample size of 3000 with 20 common items. Note that for each of the three error indices plotted (squared bias, variance, and mean squared error) the scale of the vertical axis is different for the the two plots corresponding to the 1000 sample size condition than for the two plots corresponding to the 3000 sample size condition due to the difference in the magnitude of error in the 1000 and 3000 sample size conditions.

Within each plot four points along the horizontal axis give results for 1) BILOG-MG in the form B mean 0 condition (labeled "B0"), 2) MULTILOG in the form B mean 0 condition (labeled "M0"), 3) BILOG-MG in the form B mean 1 condition (labeled "B1"), and 4) MULTILOG in the form B mean 1 condition (labeled "M1"). Five sets of these four values are given in each plot. The values within each of the five sets are connected by lines. The line with the "X" symbol gives results for concurrent estimation. The lines corresponding to a solid square, solid circle, hollow square, and hollow circle give separate estimation results for the Stocking-Lord, Haebara, Mean/Sigma, and Mean/Mean methods, respectively.

The weighted results given in Figures 1 and 3 and unweighted results given Figures 2 and 4 are similar. Except where noted, the remaining discussion applies to both the weighted and unweighted results. In addition, with a few exceptions where noted, the results described apply to both the true score equating criterion presented in Figures 1 and 2 and the ICC criterion presented in Figures 3 and 4, although the magnitude of some effects are larger for the true score equating criterion.

There are some expected results that hold for both concurrent and separate estimation for all the item parameter scaling methods. The MSE is less when the sample size is larger, and the MSE is less in the 20 common item condition than in the 10 common item condition due to increased variance in the 10 common item mean 0 condition, and increased variance and bias in the 10 common item mean 1 condition. In addition, the MSE is less in the mean 0 condition than in the mean 1 condition, with much of this due to increased bias in the mean 1 condition. The following sections discuss differences among the item parameter scaling methods, differences between concurrent and separate estimation, and differences between MULTILOG and BILOG-MG.

Item Parameter Scaling Methods

The Stocking-Lord and Haebara methods have lower MSE than the Mean/Mean and Mean/Sigma methods, and this effect is quite pronounced for the true score equating criterion. The lower MSE of the Stocking-Lord and Haebara methods is primarily due to the lower variance of these methods. The MSEs for the Stocking-Lord and Haebara methods are similar to one another and neither method has consistently lower MSE than the other. With a few exceptions, the MSE of the Mean/Mean method is less than the MSE of the Mean/Sigma method. The lower MSE of the Mean/Mean method is due to the Mean/Mean method having lower variance than the Mean/Sigma method. The biases of the Mean/Mean and Mean/Sigma methods are similar in the mean 0 condition, but the bias of the Mean/Mean method is higher in the mean 1 condition.

In the mean 0 condition item parameter scaling is not strictly necessary due to the groups taking the two forms being randomly equivalent. Figure 5 presents values of the weighted and unweighted average true score equating criterion for the mean 0 condition when no parameter scaling is used and for the four parameter scaling methods. The results for the four parameter scaling methods in Figure 5 are the same as those presented in Figures 1 and 2. The MSE when there is no parameter scaling tends to be greater than the MSE for the Stocking-Lord and Haebara methods, but less than the MSE for the Mean/Mean and Mean/Sigma methods.

Concurrent versus Separate Estimation

The relative performance of concurrent versus separate estimation interacts with whether MULTILOG or BILOG-MG is used for parameter estimation. The error in the concurrent estimates will be compared to the error in the separate estimates using the Stocking-Lord and Haebara scaling methods, since the Stocking-Lord and Haebara methods performed much better than the Mean/Mean and Mean/Sigma methods.

For BILOG-MG the concurrent estimates always result in lower MSE than the separate estimates with the exception of the unweighted ICC criterion in the mean 0, 10 common item, 3000 sample size condition, although the effect is larger in the mean 1 condition than in the mean 0 condition. In the mean 0 condition most of the difference in MSE is due to a difference in variance. In the mean 1 condition the difference in MSE is due to both a difference in variance and squared bias, but more due to a difference in squared bias.

For MULTILOG the concurrent estimates had lower MSE than the separate estimates in the mean 0 condition. This was primarily due to the concurrent estimates having lower variance in the mean 0 condition. The separate estimates using Stocking-Lord (and in some cases Haebara) scaling had lower MSE than the concurrent estimates in the mean 1 condition for the true score equating criterion. For the ICC criterion the concurrent estimates had lower MSE in the mean 1 condition.

Figures 6 and 7 contain separate values of the average weighted ICC criterion for the common and noncommon items. Figure 6 presents results for MULTILOG and Figure 7 presents results for BILOG-MG. For the concurrent estimates the MSE is lower for the common items than the noncommon items. For the separate estimates there is less difference in the MSE for the common and noncommon items than for the concurrent estimates. Thus, the difference between the concurrent and separate estimates is bigger for the common items than for the noncommon items.

Figures 8 through 11 present squared bias, variance, and MSE of the true score equating criterion across new form raw score points for the MULTILOG mean 0, MULTILOG mean 1, BILOG-MG mean 0, and BILOG-MG mean 1 conditions, respectively. The four plots present results for the 20 common item, 3000 sample size condition. The pattern of results in Figures 8 through 11 are similar to those for other conditions with different combinations of sample size and number of common items.

There is variation in the errors across score points. Squared bias tends to be much higher at lower raw score points, with the most pronounced effect in the mean 1 condition. The variance does not vary as much across score points except for the Mean/Sigma method. The pattern of errors is similar for the concurrent, Stocking-Lord, and Haebara methods, except for the BILOG-MG mean 1 condition in which concurrent estimates have much lower squared bias at low raw scores than the separate estimates. The pattern of errors for the Mean/Sigma and Mean/Mean methods differs from that of the other methods. The concurrent method has consistently lower variance over all score points, but the method with the lowest bias differs across score points. For example, in Figure 9 it can be seen that the lower average MSE of the Stocking-Lord method versus the concurrent method for the MULTILOG mean 1 condition, as reported in Figure 1, is due to the lower bias of the Stocking-Lord method at low score points.

MULTILOG versus BILOG-MG

MULTILOG and BILOG-MG tended to perform similarly, although there were some consistent differences in MSE for the two programs across conditions. The MSE using MULTILOG tends to be the same or lower than the MSE using BILOG-MG in the mean 0 condition for both concurrent and separate estimation. In the mean 1 condition BILOG-MG had lower MSE for concurrent estimation, but MULTILOG tended to have lower MSE for separate estimation.

Discussion

This paper used simulation to investigate the performance of concurrent versus separate estimation in putting item parameter estimates for two forms of a test administered in a common item equating design on the same scale. As with any simulation study considerable caution needs to be exercised in drawing conclusions due to the small number of conditions investigated. In this case, the results pertain to only the two specific forms used in this study. In addition, only two sample sizes, two levels of the number of common items, and only one level of the difference between the population distributions taking the two forms were considered.

The differences among the item parameter scaling methods used in separate estimation were much larger than the differences between concurrent estimation and the better performing scaling methods in separate estimation. The errors in the Mean/Sigma and Mean/Mean methods were substantially larger than the errors in the Stocking-Lord and Haebara methods or in concurrent estimation. The magnitude of these results suggest that item characteristic curve methods for item parameter scaling (Stocking-Lord and Haebara) should be preferred over moment methods (Mean/Mean and Mean/Sigma). The preference for characteristic curve methods is consistent with the recommendation of Kolen and Brennan (1985) based on their review of research comparing parameter scaling methods. The results also indicate that if common items are available with randomly equivalent groups it is beneficial to perform an item parameter scaling using a characteristic curve method even though it is strictly not needed.

MULTILOG and BILOG-MG tended to performed similarly, although there were some consistent patterns of differences in MSE for the two programs. With equivalent groups (mean 0 conditions) the MSE of MULTILOG and BILOG-MG were similar, except for the unweighted ICC criterion where the MSE of MULTILOG was smaller. In the case of concurrent estimates with nonequivalent groups the MSE of BILOG-MG was smaller than the MSE of MULTILOG, although in these cases MULTILOG tended to not reach convergence after 100 EM cycles. These results depend on the particular options used for these programs in this study (see Appendices A and B). The relative performance of the two programs may change when different options are used. While there were no convergence problems in the separate estimation conditions, there were some convergence problems with both MULTILOG and BILOG-MG in the concurrent estimation conditions with nonequivalent groups. Concurrent estimation puts more of a burden on the programs than separate estimation, which may result in some performance problems. This could especially be true in larger problems with more than two forms being equated simultaneously.

Except for the MULTILOG mean 1 condition, concurrent estimation produced lower errors than separate estimation. By examining the ICC criterion for common and noncommon item separately (Figures 6 and 7) it appears that at least some of the advantage of the concurrent estimates is due to their lower error on the common items, which is expected due to the common item parameter estimates being based on larger samples for the concurrent versus separate estimation. In this study the decision was made to use the form A parameter estimates for the common items when computing criteria in the separate estimation conditions. In contrast, Kim and Cohen (1998) used the average of the common item parameter estimates from the two forms. The error in the separate estimates could possibly be reduced if the form B data are used in some way to update the common item parameter estimates obtained from the form A data. Simply averaging the two estimates seems rather ad hoc (e.g., it ignores any differences that might exist in the precision of the common item parameter estimates obtained from forms A and B).

Even though concurrent estimation generally resulted in lower errors than separate estimation in this study we do not believe that the results, together with previous research on this topic, are sufficient to recommend completely avoiding separate estimation in favor of concurrent estimation. Partly, this is due to the inconsistent results observed regarding concurrent and separate estimation. In the more important case of nonequivalent groups (mean 1 condition) concurrent estimation results in lower error when using BILOG-MG, but for the true score equating criterion separate estimation results in lower error when using MULTILOG. In interpreting this finding it should be remembered that the cases in which separate estimation resulted in lower error than concurrent estimation with MULTILOG were cases in which the convergence criterion in MULTILOG was not met, although the criterion appeared close to being met. Previous research has also not shown consistent findings in favor of concurrent estimation. Some studies have concluded that concurrent estimation performed somewhat better than separate estimation (Petersen, Cook, and Stocking, 1983; Wingersky, Cook, and Eignor, 1987), while Kim and Cohen (1998), using computer programs more commonly used today, concluded that the performance of separate estimation was equal to or better than concurrent estimation.

A potential benefit of separate estimation is that having two sets of item parameter estimates can help to identify potential problems. For example, the first author has seen several cases in which close examination of the two sets of item parameter estimates for the common items obtained from separate estimation resulted in the identification of serious problems that would have remained undetected if only a single set of item parameter estimates existed for the common items. In some cases problems have been discovered by using separate estimation, computing all four item parameter scaling methods used in this paper, and looking for large discrepancies among the parameter scaling methods. Thus, we agree with the recommendation of Kolen and Brennan (1995) that it is useful for diagnostic purposes to apply multiple item parameter scaling methods when using separate estimation. Even if one wished to use concurrent estimates operationally, separate estimates could still be computed for diagnostic purposes. Separate estimates could be used to identify potential problems, and concurrent estimates used as the operational item parameter estimates.

An important factor to consider when interpreting the results of this study, and other studies of this type, is that the data were simulated from the same model used for item parameter estimation. With real data the simple unidimensional model used in this paper would probably be misspecified to some extent, and this could affect the relative performance of separate versus concurrent estimation. Using separate estimation to obtain item parameter estimates for separate intact forms that have a proper content balance may reduce the effects of model misspecification in contrast to estimating a group of items together with concurrent estimation that do not represent the proportional content balance of an intact form. An important further research question is how well separate and concurrent estimation perform when the model is misspecified to some degree, as would occur with real data.

An alternative procedure not studied in this paper that combines features of concurrent and separate estimation is to estimate item parameters for one form and then estimate the parameters in the other form with the common item parameters fixed at their estimated values using the first form. This procedure is called the fixed parameter or item anchoring method. Results in this paper suggested that a factor favoring concurrent estimation over separate estimation is the larger sample size used to estimate item parameters for the common items. This advantage of concurrent estimation would also hold over the fixed parameter method.

Appendix A

This appendix gives the MULTILOG control files used in the separate estimation condition and the concurrent estimation conditions with 10 and 20 common items. The statement "NE=3000" (separate) or "NE=6000" (concurrent) in the PROBLEM command used when the sample size was 3000 was changed to "NE=1000" (separate) or "NE=2000" (concurrent) when the sample size was 1000. For the concurrent estimation condition the statement "NG=2" in the PROBLEM commend (used for the mean 1 condition) was changed to "NG=1" in the mean 0 condition.

Separate Estimation

```
Separate estimation
>PROBLEM RA IN NI=60 NE=3000 NG=1;
>TEST AL L3;
>PRIORS AL CJ PARAMS=(-1.38, 0.5);
>PRIORS AL AJ PARAMS=(1.0, 1.0);
>TGROUPS NU=40,
QP=(-4.0000, -3.7950, -3.5900, -3.3850, -3.1790,
-2.9740, -2.7690, -2.5640, -2.3590, -2.1540,
-1.9490, -1.7440, -1.5380, -1.3330, -1.1280,
-0.9231, -0.7179, -0.5128, -0.3077, -0.1026,
0.1026, 0.3077, 0.5128, 0.7179, 0.9231,
1.1280, 1.3330, 1.5380, 1.7440, 1.9490,
2.1540, 2.3590, 2.5640, 2.7690, 2.9740,
3.1790, 3.3850, 3.5900, 3.7950, 4.0000);
>SAVE;
>ESTIMATE NC=100;
>END;
2
01
111111111111111111111111111111111111111111111111111111111111
N
(60A1)
```

Concurrent Estimation with 20 Common Items

```
Concurrent estimation with 20 common items
>PROBLEM RA IN NI=100 NE=6000 NG=2;
>TEST AL L3;
>PRIORS AL CJ PARAMS=(-1.38, 0.5);
>PRIORS AL AJ PARAMS=(1.0, 1.0);
>TGROUPS NU=40,
QP=(-4.0000, -3.7950, -3.5900, -3.3850, -3.1790,
-2.9740, -2.7690, -2.5640, -2.3590, -2.1540,
-1.9490, -1.7440, -1.5380, -1.3330, -1.1280,
-0.9231, -0.7179, -0.5128, -0.3077, -0.1026,
0.1026, 0.3077, 0.5128, 0.7179, 0.9231,
1.1280, 1.3330, 1.5380, 1.7440, 1.9490,
2.1540, 2.3590, 2.5640, 2.7690, 2.9740,
3.1790, 3.3850, 3.5900, 3.7950, 4.0000);
>SAVE;
```

```
>ESTIMATE NC=100;
>END;
3
019
11111111111111111111111111111111111111111111111111111111111111111111111111111111
111111111111111111111
Y
9
```

Concurrent Estimation with 10 Common Items

```
Concurrent estimation with 10 common items
>PROBLEM RA IN NI=110 NE=6000 NG=2;
>TEST AL L3;
>PRIORS AL CJ PARAMS=(-1.38, 0.5);
>PRIORS AL AJ PARAMS=(1.0, 1.0);
>TGROUPS NU=40,
QP=(-4.0000, -3.7950, -3.5900, -3.3850, -3.1790,
-2.9740, -2.7690, -2.5640, -2.3590, -2.1540,
-1.9490, -1.7440, -1.5380, -1.3330, -1.1280,
-0.9231, -0.7179, -0.5128, -0.3077, -0.1026,
0.1026, 0.3077, 0.5128, 0.7179, 0.9231,
1.1280, 1.3330, 1.5380, 1.7440, 1.9490,
2.1540, 2.3590, 2.5640, 2.7690, 2.9740,
3.1790, 3.3850, 3.5900, 3.7950, 4.0000);
>SAVE;
>ESTIMATE NC=100;
>END;
3
019
11111111111111111111111111111111111111111111111111111111111111111111111111111111
1111111111111111111111111111111
Y
9
```

Appendix B

This appendix gives the BILOG-MG control files used in the separate estimation condition, the concurrent estimation condition with 20 common items and equivalent groups and the concurrent estimation condition with 10 common items and 2 groups. The statements "SAMPLE=3000" and "SAMPLE=6000" (used when the sample size was 3000) in the INPUT command where changed to "SAMPLE=1000" and "SAMPLE=2000" when the sample size was 1000.

Separate Estimation

```
>COM separate estimation
>GLO DFN='datafile',NPAR=3, NTES=1, SAVE;
>SAV PAR='parameterfile';
>LENGTH NIT=60;
>INPUT NTO=60, TYP=1, NAL=4, NIDCH=2, SAMPLE=3000;
>ITEMS;
>TES INUM=(1(1)60);
(2A1,T6,60A1)
>CAL NQPT=40, CYC=20, NEW=15, CASE=3;
>SCO;
```

Concurrent Estimation with 20 Common Items and 1 Group

```
>COM concurrent estimation with 20 common items and 1 group
>GLO DFN='datafile',NPAR=3, NTES=1, SAVE;
>SAV PAR='parameterfile';
>LENGTH NIT=100;
>INPUT NTO=100, TYP=1, NAL=4, NIDCH=2, NFO=2, SAMPLE=6000;
>ITEMS;
>TES INUM=(1(1)100);
>FORM1 LEN=60, INU=(1(1)60);
>FORM2 LEN=60, INU=(41(1)100);
(2A1,I1,T6,60A1)
>CAL NQPT=40, CYC=20, NEW=15, CASE=3;
>SCO;
```

Concurrent Estimation with 10 Common Items and 2 Groups

```
>COM concurrent estimation with 10 common items and 2 groups
>GLO DFN='datafile',NPAR=3, NTES=1, SAVE;
>SAV PAR='parameterfile';
>LENGTH NIT=110;
>INPUT NTO=110, TYP=1, NAL=4, NIDCH=2, NFO=2, NGRO=2, SAMPLE=6000;
>ITEMS;
>TES INUM=(1(1)110);
>FORM1 LEN=60, INU=(1(1)60);
>FORM2 LEN=60, INU=(51(1)110);
>GROUP1 GNAME='base', LEN=60, INU=(1(1)60);
>GROUP2 GNAME='comp', LEN=60, INU=(51(1)110);
(2A1,2I1,T6,60A1)
>CAL NQPT=40, CYC=20, NEW=15, CASE=3;
>SCO;
```

References

ACT, Inc. (1997). *ACT Assessment Technical Manual.* Iowa City, IA: Author.

Kim, S., & Cohen, A. S. (1998). A comparison of linking and concurrent calibration under item response theory. *Applied Psychological Measurement, 22,* 131-143.

Kolen, M. J., & Brennan, R. L. (1995). *Test equating methods and practices.* New York: Springer-Verlag.

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

Petersen, N. S., Cook, L. L., & Stocking, M. L. (1983). IRT versus conventional equating methods: A comparative study of scale stability. *Journal of Educational Statistics, 8,* 137-156.

Mislevy, R. J., & Bock, R. D. (1990). *BILOG 3* (2nd ed.) [Computer program]. Mooresville IN: Scientific Software International.

Thissen, D. (1991). *Multilog user's guide: Multiple, categorical item analysis and test scoring using item response theory* [Computer program]. Chicago: Scientific Software International.

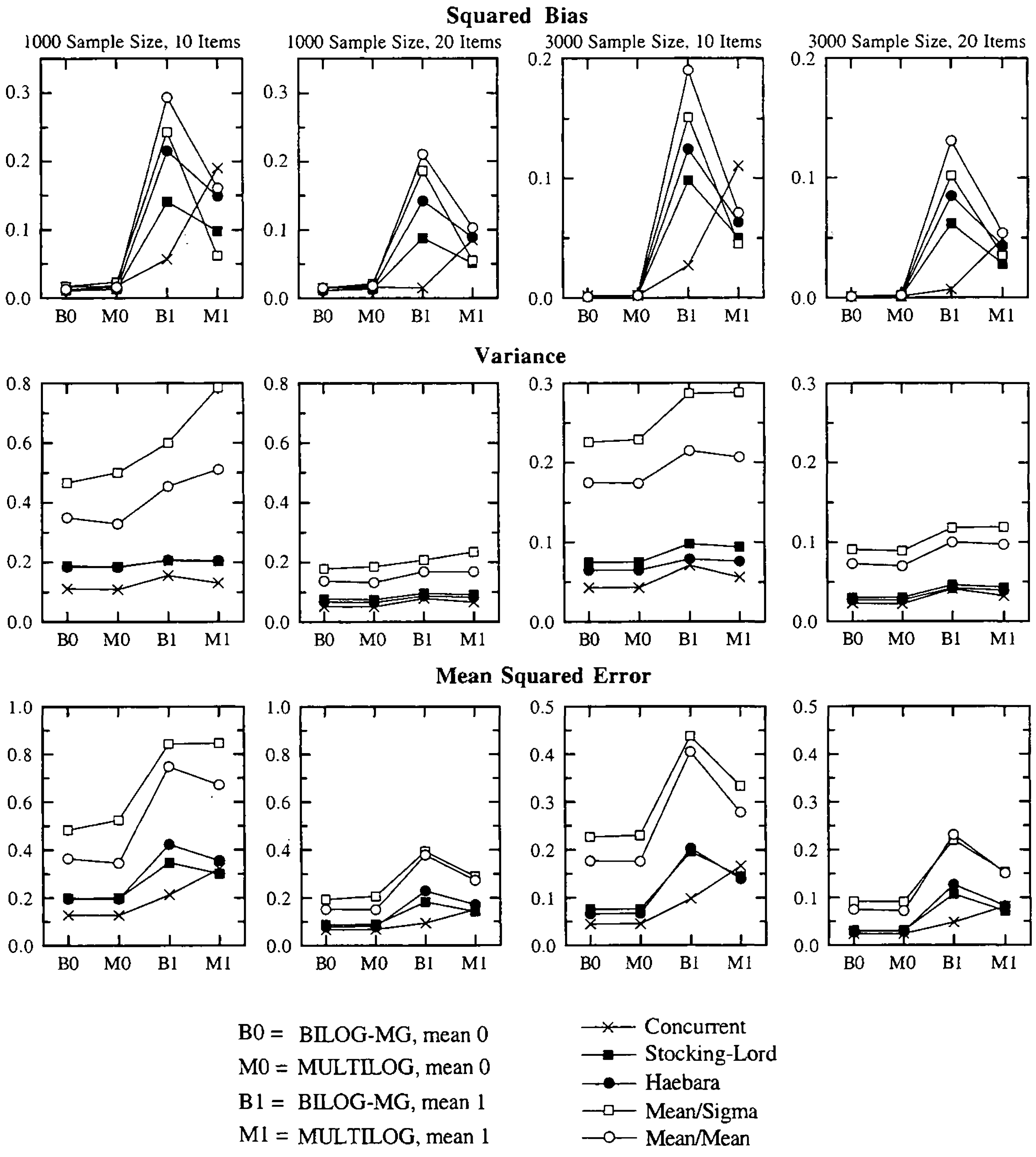
Wingersky, M. S., Barton, M. A., & Lord, R. M. (1982). *LOGIST user's guide* [Computer program]. Princeton NJ: Educational Testing Service.

Wingersky, M. S., Cook, L. L., & Eignor, D. R. (1987). *Specifying the characteristics of linking items used for item response theory item calibration.* ETS Research Report 87-24. Princeton NJ: Educational Testing Service.

Zimowski, M. F., Muraki, E., Mislevy, R. J., & Bock, R. D. (1996). *BILOG-MG: Multiple-group IRT Analysis and Test Maintenance for Binary Items* [Computer program]. Chicago: Scientific Software International.

Table 1. Population Item Parameters Used for Simulations.

	Parameters				Parameters		
Item	a	b	c	Item	a	b	c
1	0.642	-2.522	0.187	51	0.957	0.192	0.194
2	0.806	-1.902	0.149	52	1.269	0.683	0.150
3	0.956	-1.351	0.108	53	1.664	1.017	0.162
4	0.972	-1.092	0.142	54	1.511	1.393	0.123
5	1.045	-0.234	0.373	55	0.561	-1.865	0.240
6	0.834	-0.317	0.135	56	0.728	-0.678	0.244
7	0.614	0.037	0.172	57	1.665	-0.036	0.109
8	0.796	0.268	0.101	58	1.401	0.117	0.057
9	1.171	-0.571	0.192	59	1.391	0.031	0.181
10	1.514	0.317	0.312	60	1.259	0.259	0.229
11	0.842	0.295	0.211	61	0.804	-2.283	0.192
12	1.754	0.778	0.123	62	0.734	-1.475	0.233
13	0.839	1.514	0.170	63	1.523	-0.995	0.175
14	0.998	1.744	0.057	64	0.720	-1.068	0.128
15	0.727	1.951	0.194	65	0.892	-0.334	0.211
16	0.892	-1.152	0.238	66	1.217	-0.290	0.138
17	0.789	-0.526	0.115	67	0.891	0.157	0.162
18	1.604	1.104	0.475	68	0.972	0.256	0.126
19	0.722	0.961	0.151	69	1.206	-0.463	0.269
20	1.549	1.314	0.197	70	1.354	0.122	0.211
21	0.700	-2.198	0.184	71	0.935	-0.061	0.086
22	0.799	-1.621	0.141	72	1.438	0.692	0.209
23	1.022	-0.761	0.439	73	1.613	0.686	0.096
24	0.860	-1.179	0.131	74	1.199	1.097	0.032
25	1.248	-0.610	0.145	75	0.786	-1.132	0.226
26	0.896	-0.291	0.082	76	1.041	0.131	0.150
27	0.679	0.067	0.161	77	1.285	0.170	0.077
28	0.996	0.706	0.210	78	1.219	0.605	0.128
29	0.420	-2.713	0.171	79	1.473	1.668	0.187
30	0.977	0.213	0.280	80	1.334	0.530	0.075
31	1.257	0.116	0.209	81	0.965	-1.862	0.152
32	0.984	0.273	0.121	82	0.710	-1.589	0.138
33	1.174	0.840	0.091	83	0.523	-1.754	0.149
34	1.601	0.745	0.043	84	1.134	-0.604	0.181
35	1.876	1.485	0.177	85	0.709	-0.680	0.064
36	0.620	-1.208	0.191	86	0.496	-0.443	0.142
37	0.994	0.189	0.242	87	0.979	0.181	0.124
38	1.246	0.345	0.187	88	0.970	0.351	0.151
39	1.175	0.962	0.100	89	0.524	-2.265	0.220
40	1.715	1.592	0.096	90	0.944	-0.084	0.432
41	0.769	-1.944	0.161	91	0.833	0.137	0.202
42	0.934	-1.348	0.174	92	1.127	0.478	0.199
43	0.496	-1.348	0.328	93	0.893	0.496	0.100
44	0.888	-0.859	0.199	94	1.215	0.867	0.076
45	0.953	-0.190	0.212	95	1.079	-0.486	0.264
46	1.022	-0.116	0.158	96	0.932	0.450	0.259
47	1.012	0.421	0.288	97	1.141	0.344	0.071
48	1.605	1.377	0.120	98	1.068	0.893	0.153
49	1.009	-1.126	0.133	99	1.217	1.487	0.069
50	1.310	-0.067	0.141	100	1.310	1.186	0.153


Squared Bias

1000 Sample Size, 10 Items

1000 Sample Size, 20 Items

3000 Sample Size, 10 Items

3000 Sample Size, 20 Items

Variance

Mean Squared Error

Figure 1. Average Squared Bias, Variance, and MSE Across Score Points for the Weighted True Score Equating Criterion.

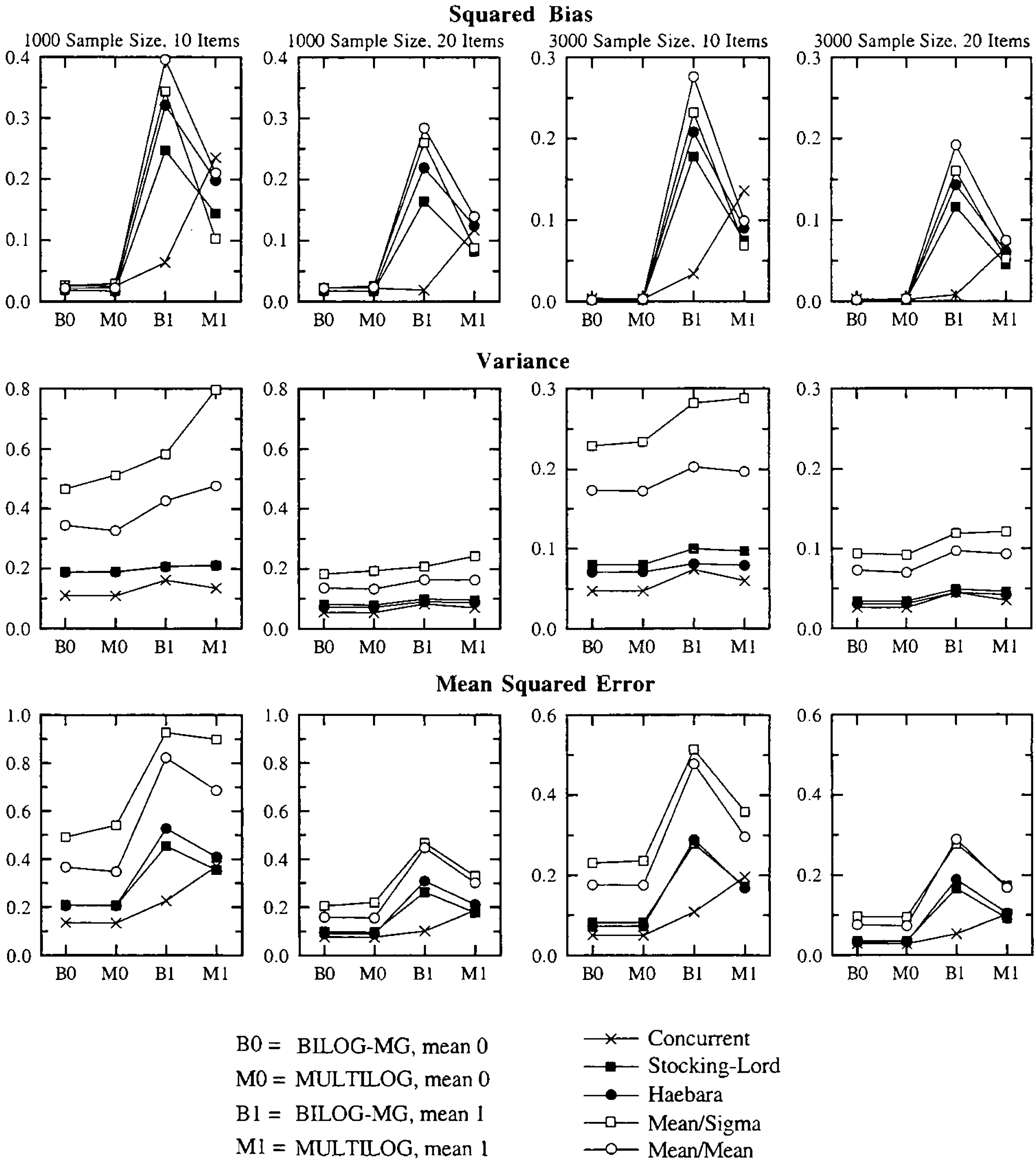


Figure 2. Average Squared Bias, Variance, and MSE Across Score Points for the Unweighted True Score Equating Criterion.

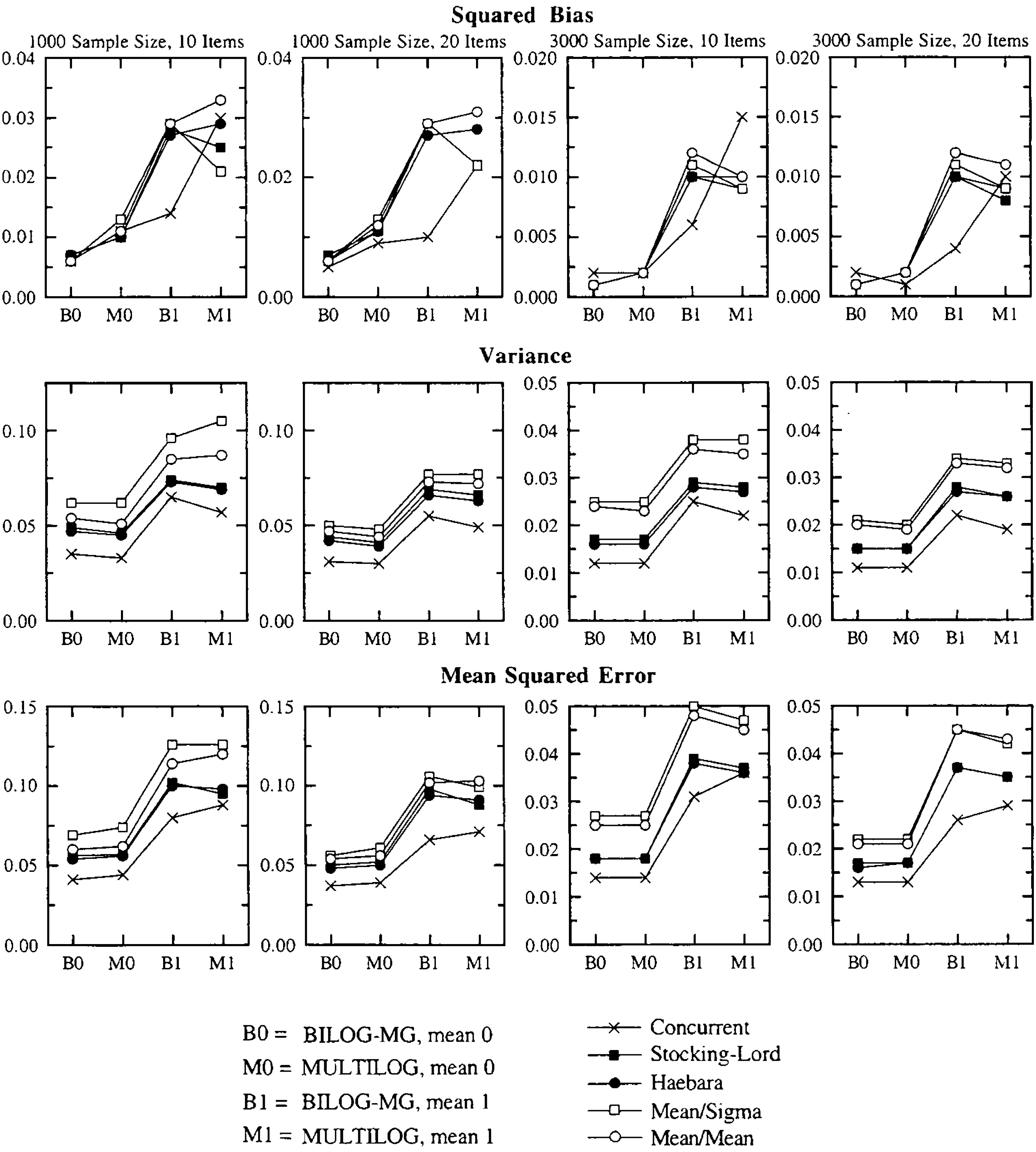


Figure 3. Average Squared Bias, Variance, and MSE Across Items for the Weighted ICC Criterion.

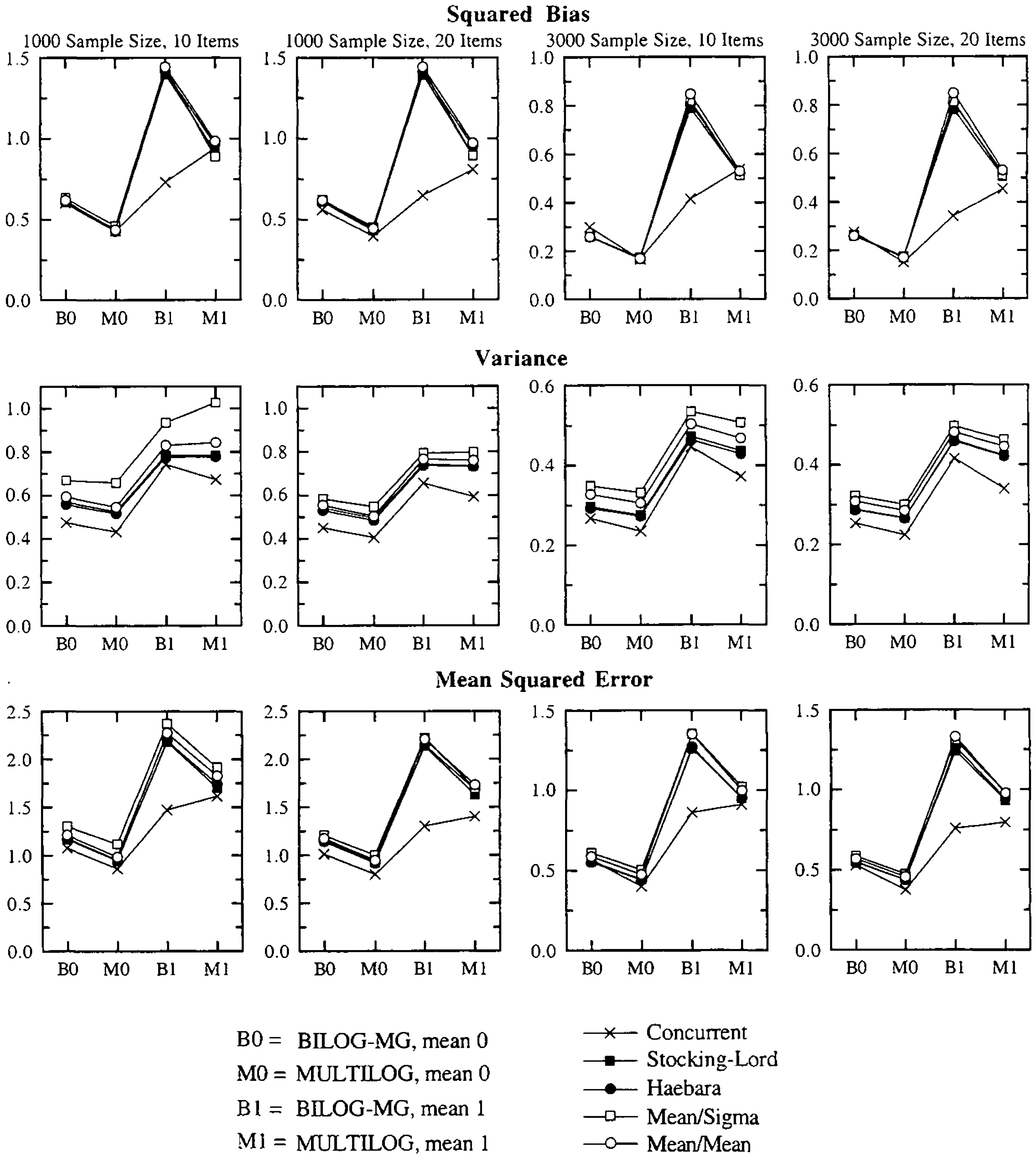


Figure 4. Average Squared Bias, Variance, and MSE Across Items for the Unweighted ICC Criterion.

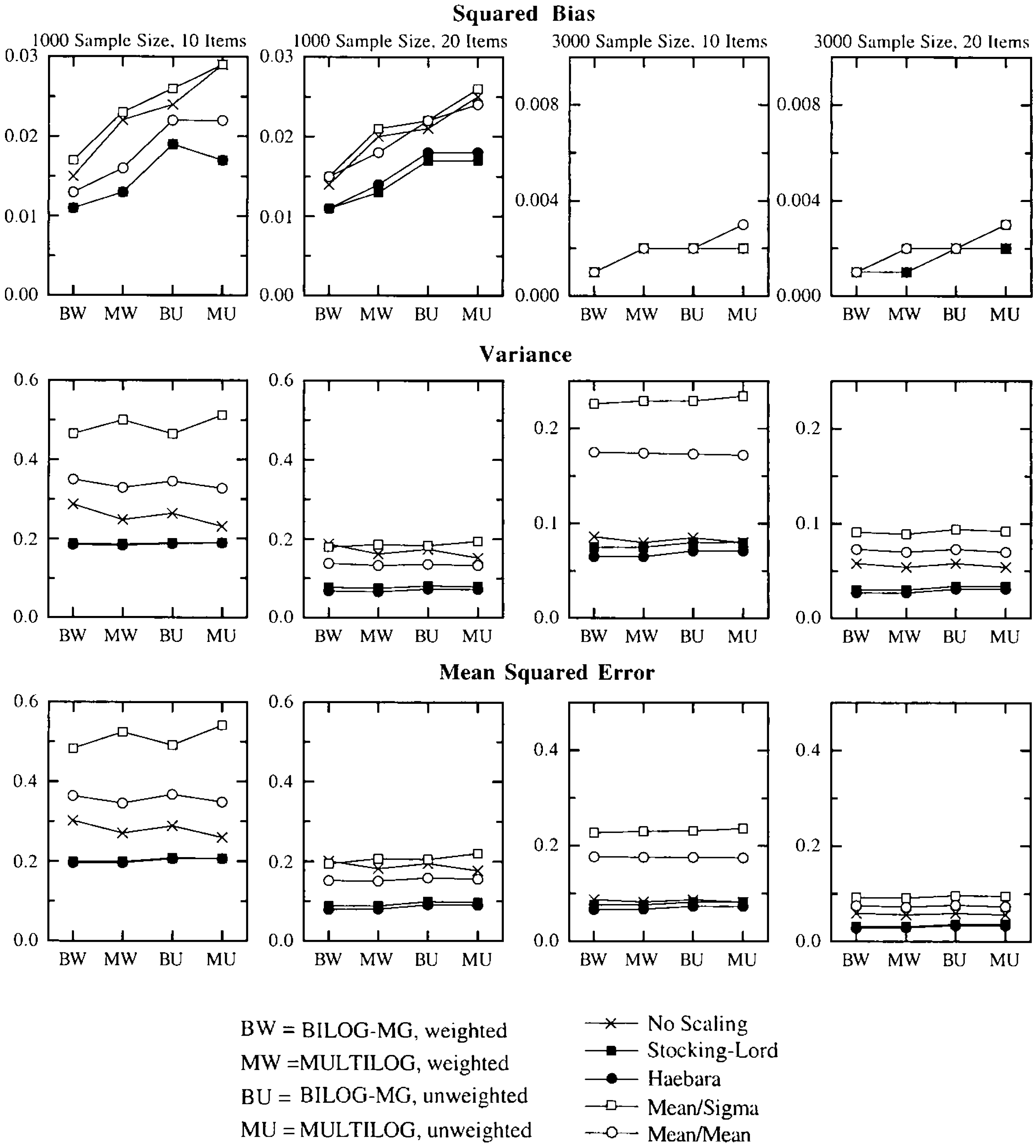


Figure 5. Average Squared Bias, Variance, and MSE Across Score Points for the True Score Equating Criterion in the Mean 0 Condition.

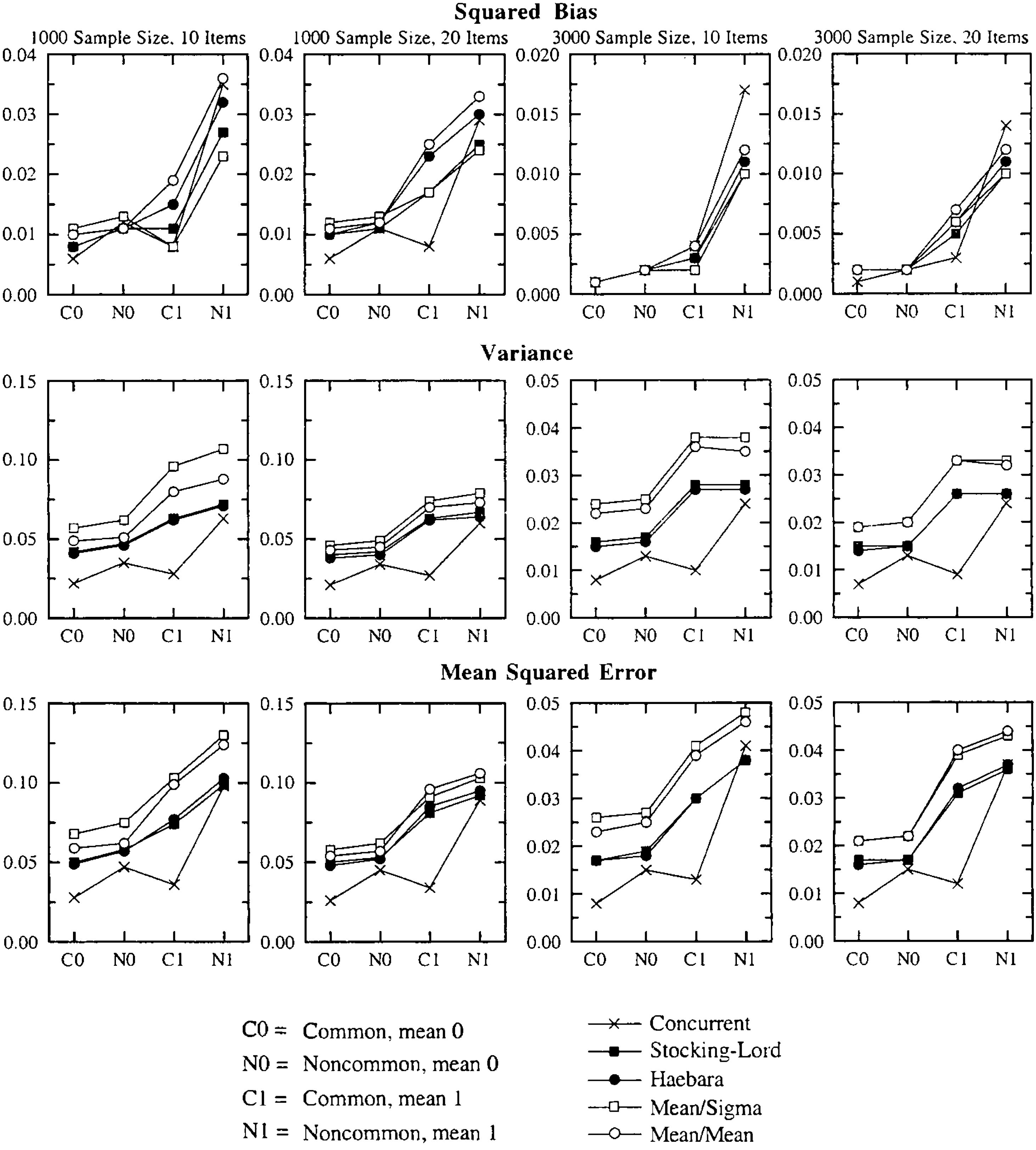


Figure 6. Average Squared Bias, Variance, and MSE Across Common and Noncommon Items for the Weighted ICC Criterion (MULTILOG).

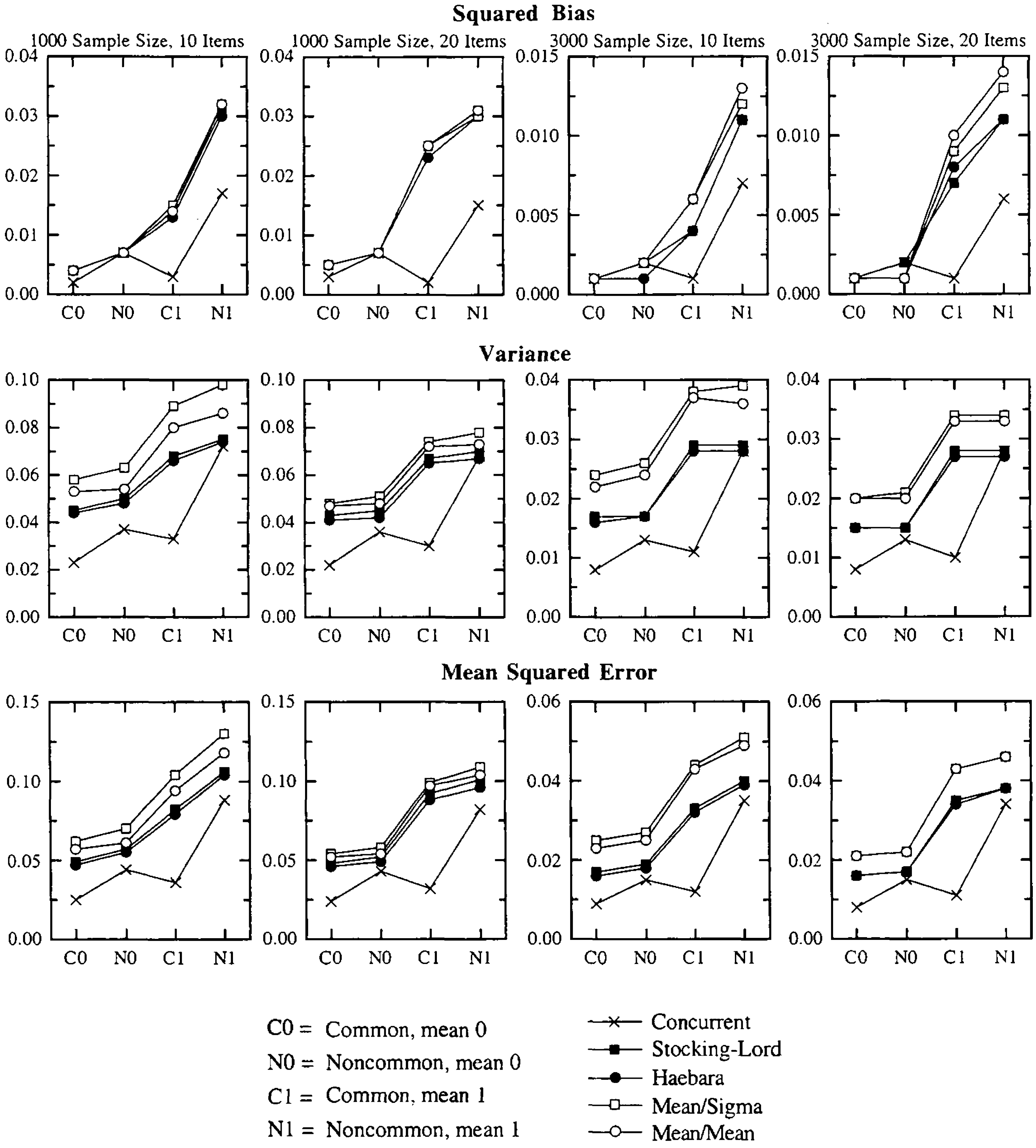


Figure 7. Average Squared Bias, Variance, and MSE Across Common and Noncommon Items for the Weighted ICC Criterion (BILOG-MG).

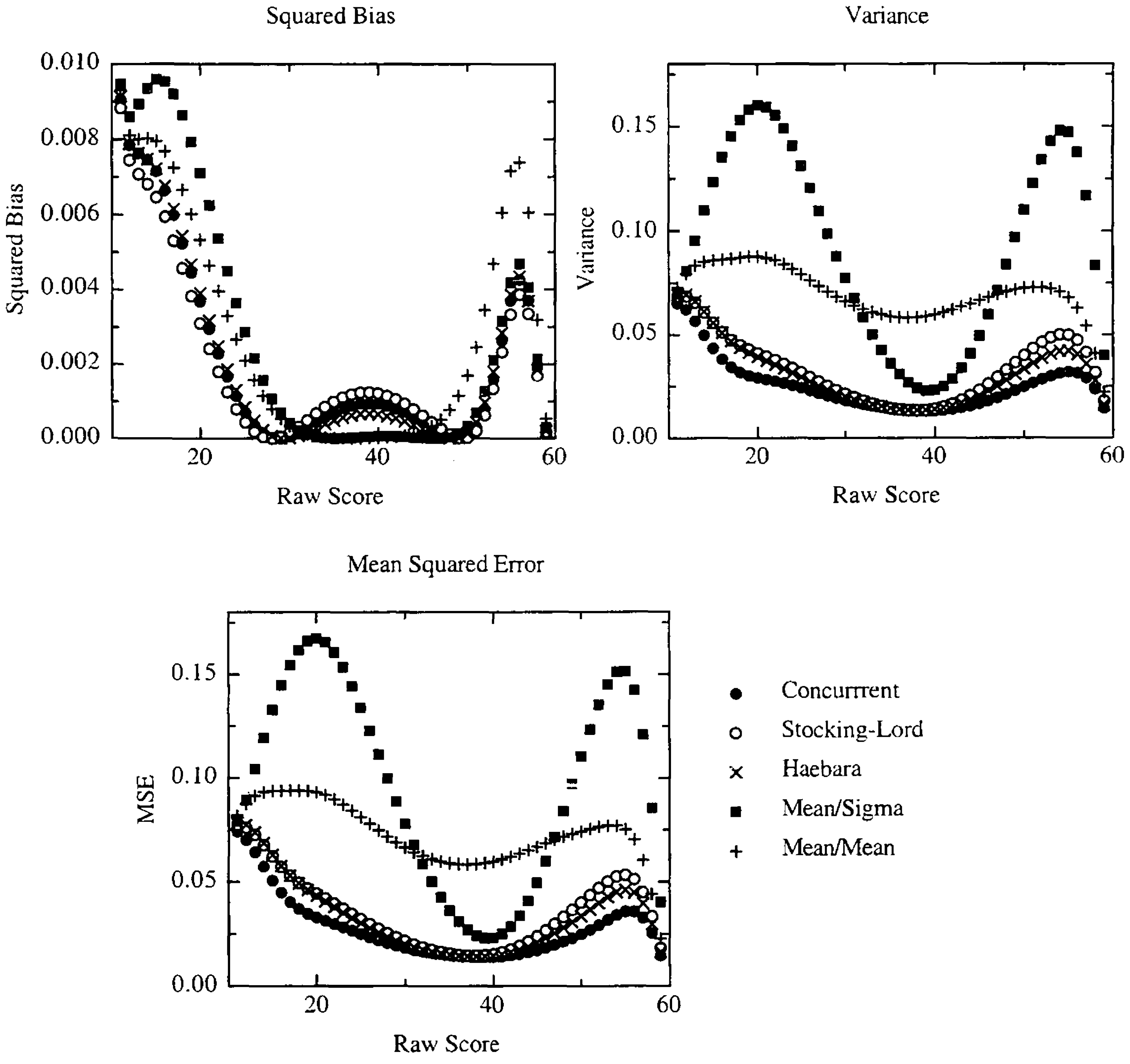


Figure 8. True Score Equating Criterion for MULTILOG Mean 0
(20 Common Items, 3000 Sample Size).

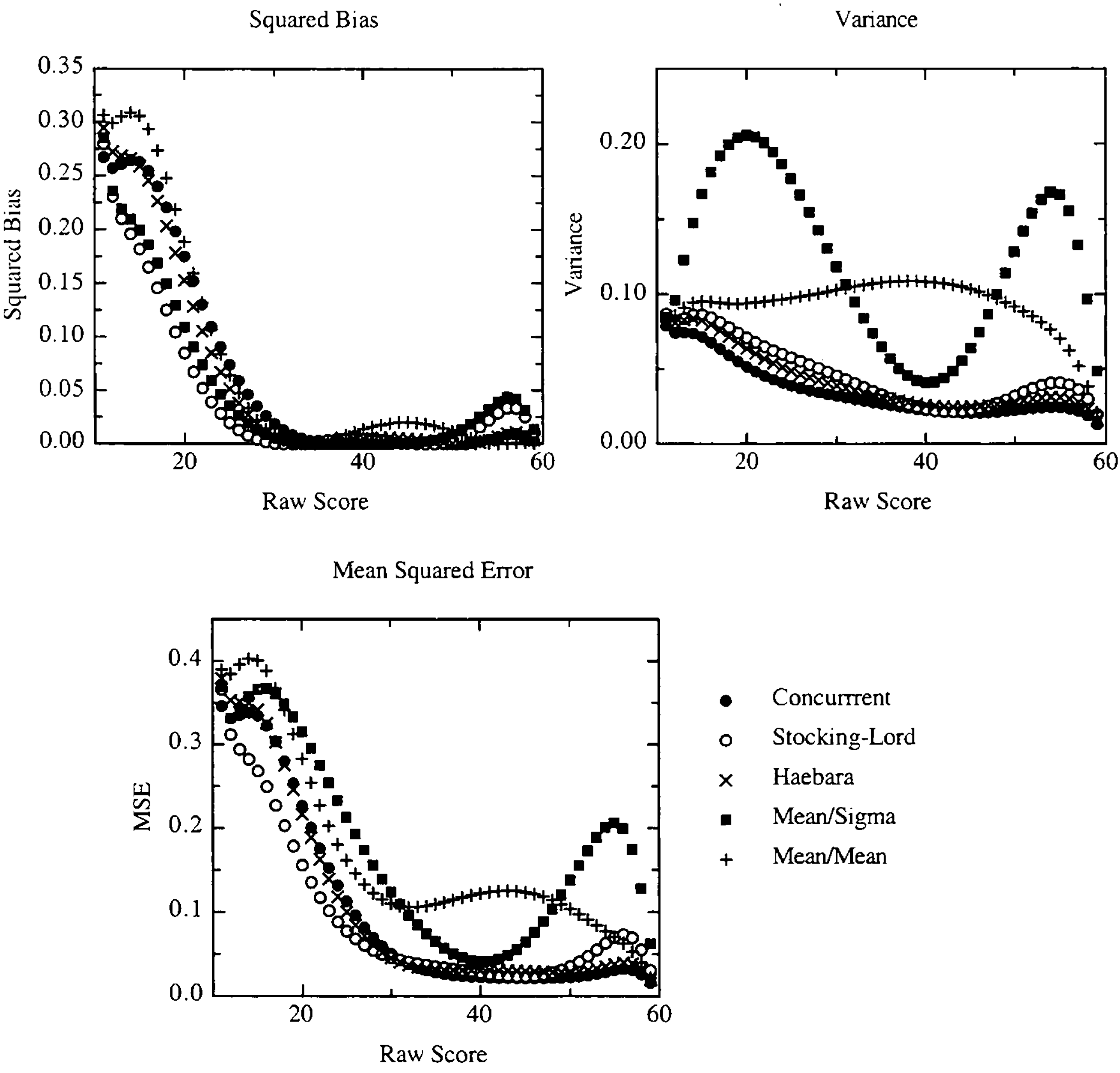


Figure 9. True Score Equating Criterion for MULTILOG Mean 1.
(20 Common Items, 3000 Sample Size)

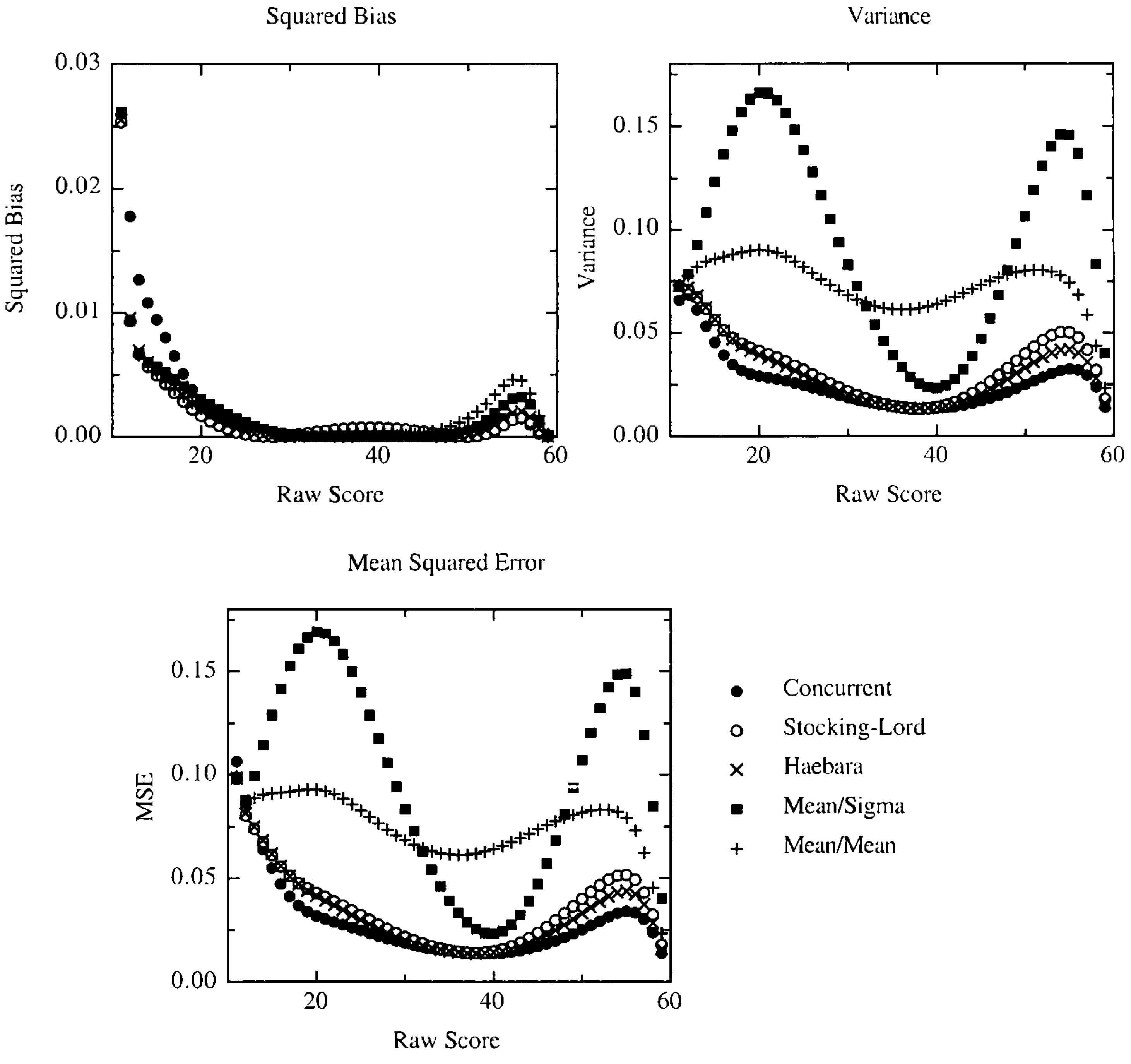


Figure 10. True Score Equating Criterion for BILOG-MG Mean 0
(20 Common Items, 3000 Sample Size).

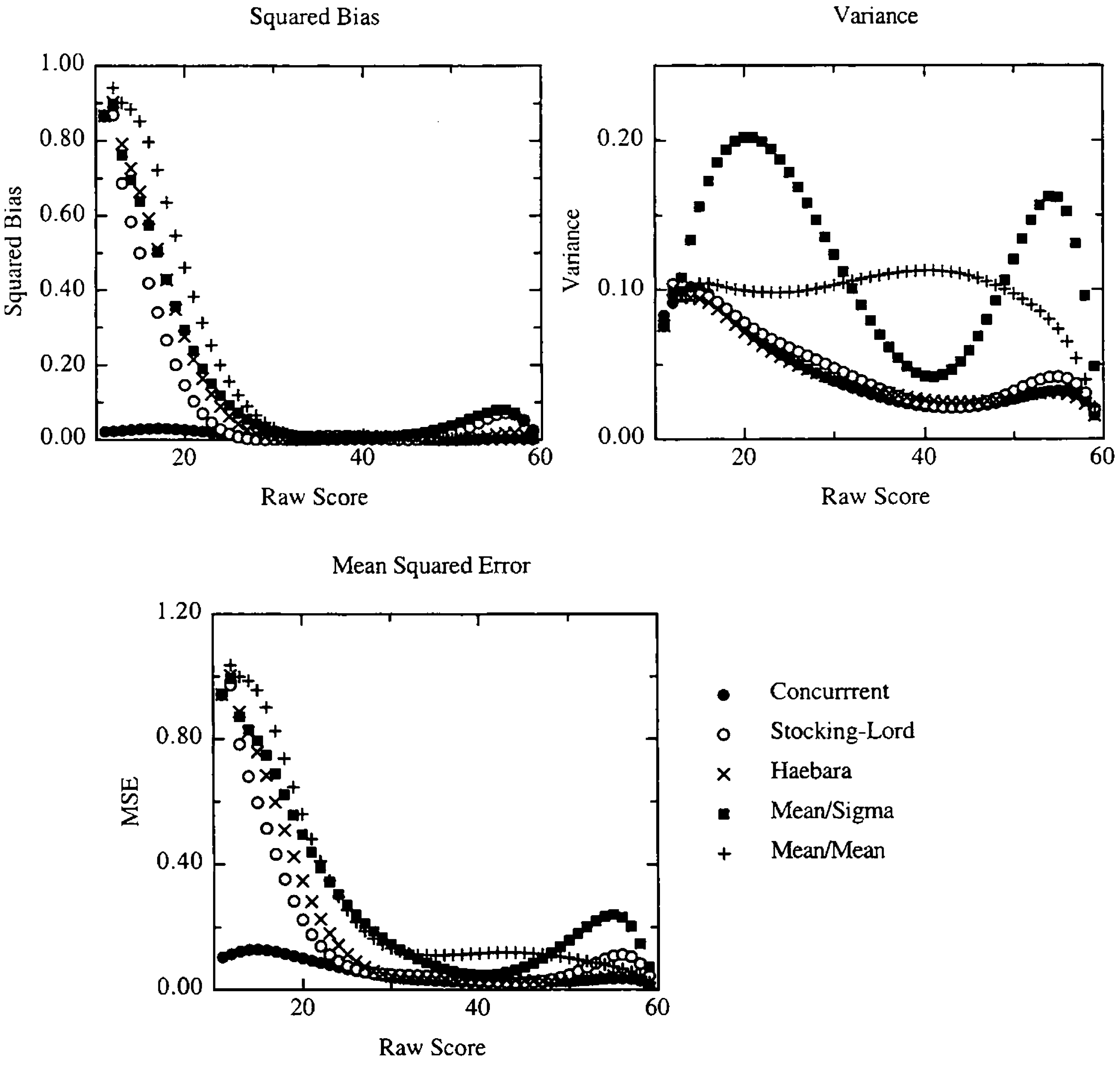


Figure 11. True Score Equating Criterion for BILOG-MG Mean 1
(20 Common Items, 3000 Sample Size).