# Calibrating CAT Pools and Online Pretest Items Using Marginal Maximum Likelihood Methods

Mary Pommerich
Daniel O. Segall

*Defense Manpower Data Center*

Paper presented at the Annual Meeting of the National Council on Measurement in Education, Chicago, IL, April 2003.

## Calibrating CAT Pools and Online Pretest Items Using Marginal Maximum Likelihood Methods

The research in this paper was conducted as part of an ongoing large-scale simulation study to evaluate methods of calibrating pretest items for CAT pools. The simulation study was designed to mimic the operational CAT-ASVAB testing program, whereby a single pretest item is embedded, or seeded, into the administration of an operational CAT. The overriding goal of the research is to select a calibration method that will best represent the data and maintain a consistent scale over time, as new calibrations are conducted, pretest items are formed into new pools, and operational pools are replaced with the new pools.

A series of papers accompany this paper. The history of the CAT-ASVAB testing program is discussed in Nicewander (2003), along with relevant issues in maintaining scale consistency for the CAT-ASVAB. The design used in the simulation study and a rationale for the design is discussed in Thomasson (2003), along with the procedures used to simulate items that do not conform to a 3PL model. MCMC methods of calibrating the pretest items are evaluated in Segall (2003). Nonparametric and adjusted marginal maximum likelihood methods of calibrating the pretest items are evaluated in Krass and Williams (2003). This paper evaluates MML methods of calibrating the pretest items using the 3PL model.

The goal of the study is to identify the best possible calibration method for new pretest items so that the true underlying parameters are recovered as closely as possible. The calibration problem is made difficult by the nature of the CAT data. Namely, examinees take a small percentage of differing operational CAT items and only one of a set of pretest items, creating a sparse matrix of item responses to be analyzed. Each operational item has a variable number of responses, some with very few responses, some with large numbers of responses. The operational items are typically administered to examinees within a limited range of ability. The pretest items are administered randomly to fixed numbers of examinees, but each examinee takes only one pretest item. This is contrary to a typical calibration for a fixed-form test or pretest items, whereby a fixed number of examinees take a fixed number of items.

In this study, pretest and operational CAT items are simultaneously calibrated and placed on the scale of the operational parameters from one CAT pool that is designated as the anchor CAT pool. Operationally, the anchor CAT pool is infrequently administered, so the parameters are assumed to have little drift over time. This assumption is warranted by our belief that item exposure is the most common cause of parameters displaying drift over time. In using the anchor pool to fix the scale, we assume that other sources of drift will not occur over time in the anchor pool. As our intention is to use the anchor pool to maintain continuity of scale as old pools are replaced with new pools in the operational administration, it is necessary to make this assumption. The re-estimation of operational CAT items, in addition to the estimation of the pretest items, should help account for any parameter drift that may occur in the operational CAT pools that are not fixed (i.e., all pools but the anchor CAT pool).

**The Simulation Design**

A primary objective of the simulation study is to evaluate the performance of the MML 3PL-based calibration methods when some of the items do not fit a 3PL model. In practice, we believe that not all items conform to a 3PL model. For purposes of evaluating the strengths and limitations of the calibration methods under conditions specific to our operational CAT program, the simulation was conducted in stages. In the first stage, a conventional test was simulated, where 10,000 examinees responded to a fixed number of items, and all items were generated using a 3PL model. Examinees were simulated from N(0,1), N(-1,1), and N(1,1) distributions. The first stage simulation provided an optimal calibration situation, which allowed us to minimize estimation error as much as possible, in order to identify important issues in calibrating to maintain a given scale. In the second stage, an operational CAT was simulated, where all items were generated using a 3PL model. Examinees were simulated from N(0,1), N(-1,1.2), and N(1,0.8) distributions. The second stage simulation allowed us to identify important issues in calibrating to maintain a given scale, under the added complications of a large sparse data matrix, varying sample sizes for some items, and a restricted range of ability for some items. Again, all items were generated using a 3PL model, so that results would not be confounded by fitting an inappropriate model to the items.

In the first two stages of the simulation the true item parameters were known. This enabled us to evaluate the strengths and weaknesses of the calibrations, and to adjust the calibration procedures accordingly. This paper reports results only from the first two stages. A third simulation will be conducted in the future, which is designed to allow us to evaluate the calibration methods under the most realistic conditions. For the third stage simulation, an independent party will generate realistic CAT data, and the true parameters will be unknown to us. Thus, we must implement the calibration procedures as we would operationally, without being able to make any adjustments to the procedures. In addition, some items will be generated using a 3PL model, while others will be generated using a non-parametric model. The percentage of 3PL and non-parametric items will be varied according to what is observed in real data. Examinees from different ability distributions will be generated, again according to what is observed in real data. The third stage simulation will allow us to fully evaluate the calibration methods under the most realistic conditions, when a percentage of items are not represented by a 3PL model. In the simulation, new CAT pools will be developed from the pretest items and administered operationally, replacing older operational pools. Scale drift will be evaluated as pools are updated.

The second stage simulations were designed to represent the operational CAT program and the conditions under which pretest items would be calibrated in practice. Operationally, pretest items are administered in groups of 100 items. One pretest item is randomly selected from the set of 100 and administered to an examinee during an operational session. One of four operational CAT pools is randomly administered to an examinee. Pools 1-3 are administered about 98% of the time, while Pool 4 is administered about 2% of the time. Parameters for Pools 1-3 and the pretest items will be simultaneously calibrated, as discussed above. Pool 4 will be used as the anchor CAT pool, and parameters for Pools 1-3 and the pretest items will be placed on the scale of the anchor CAT pool. Operationally, the calibrations will be conducted when the pretest items have been administered a sufficient number of times.

The test administration design for the second stage simulations is summarized in Table 1. Each examinee was administered a 15-item CAT (using Pool 1, Pool 2, Pool 3, or Pool 4), and one pretest item. Pool 1 contained 94 items, Pools 2-4 each contained 137 items, and the pretest set contained 100 items, for a total of 605 items, of which an examinee took 16. CAT Pools 1-3 were administered to a total of 40,000 examinees each (400 per pretest item per pool). Pool 4 was administered to a total of 2400 examinees (24 per pretest item). Each pretest item was administered to a total of 1224 examinees, resulting in a total of 122,400 examinees

Table 1. Test Administration Design for the CAT Simulations.

| Pretest Item | Operational CAT Pool | Number of Examinees |
|---|---|---|
| 1 | 1 | 400 |
| | 2 | 400 |
| | 3 | 400 |
| | 4 | 24 |
| | Total | 1224 |
| 2 | 1 | 400 |
| | 2 | 400 |
| | 3 | 400 |
| | 4 | 24 |
| | Total | 1224 |
| . | . | . |
| . | . | . |
| . | . | . |
| 100 | 1 | 400 |
| | 2 | 400 |
| | 3 | 400 |
| | 4 | 24 |
| | Total | 1224 |
| | | |
| All Items | 1 | 40000 |
| | 2 | 40000 |
| | 3 | 40000 |
| | 4 | 2400 |
| | Total | 122400 |

**MML Software**

A variety of commercial software programs exist that use marginal maximum likelihood (MML) methods of estimating parameters. Three popular commercial programs, Bilog-MG (Zimowski, Muraki, Mislevy, and Bock, 2003), Parscale (Muraki and Bock, 2003), and Multilog (Thissen, 2003) have recently been ported to a Windows platform, and the programs are documented together in one reference manual (du Toit, 2003). The calibration programs are designed for a

variety of purposes, and have differing capabilities. This paper focuses on Bilog-MG, which is one of the most popular 3PL calibration programs available.

Our interest was in evaluating calibration results from Bilog-MG when attempting to fix the scale of the estimated parameters to a known scale. A new feature in Bilog-MG for Windows is that selected item parameters may be fixed at pre-defined starting values and the rescaling of the latent distribution that is typically done in Bilog-MG may be suppressed. In theory, these features may be used to place parameter estimates for non-fixed items on the scale of the fixed items. In the case of our operational CAT, the items from the anchor pool (Pool 4) were treated as fixed, and the parameters were estimated for the pretest items and for Pools 1-3. Results for conventional data were evaluated first, followed by results for CAT data, so that problems due to the fixing of the scale could be identified without the added complication of sparse data matrices.

**Evaluation of Conventional 3PL Data**

The conventional data were simulated using parameters that mimicked parameters from one of the operational CAT pools. Responses were generated for 10,000 examinees taking a total of 94 items. The results were replicated for examinees drawn from three different ability distributions: N(0,1), N(-1,1), and N(1,1). Parameters for 20 of the 94 items (21%) were treated as anchor items and fixed at their known value, while the parameters for the remaining 74 items were estimated, with the intent that the parameters for the 74 "non-fixed" items would be on the same scale as the parameters for the 20 "fixed items". This enabled us to evaluate the capability of Bilog-MG to fix the scale. The percentage of fixed items used here is comparable to the percentage of fixed items that we would have for our operational CAT. Note that the "NOADJUST" option in the "CALIB" command was used in every calibration where item parameters were fixed to suppress the rescaling of the item parameters, and no Phase 3 analyses were conducted. Thus, the final parameters for the fixed items were always equal to their starting values.

Figure 1 shows the recovery of the b parameters for the non-fixed items when Bilog-MG was used to fix the parameters for 20 items and estimate the parameters for the remaining 74 items, and the data were generated from a N(0,1) distribution. Parameter recovery, as expected, was quite good. Figure 2 shows the recovery of the b parameters for the non-fixed items when the same procedure was applied to data generated from a N(+1,1) distribution. In this case, parameter recovery was not as clean. There appeared to be a slight systematic bias in the parameter estimates; the estimated parameters were somewhat underestimated. Figure 3 shows the recovery of the b parameters for the non-fixed items when the process was applied to data generated from a N(-1,1) distribution. Again, there appeared to be a slight systematic bias in the parameter estimates. In this case, the estimated parameters were somewhat inflated.

Figures 4 and 5 show the estimated posterior distributions from Bilog-MG for the data simulated from a N(+1,1) and N(-1,1) distribution, respectively, relative to the population distribution from which the data were generated. Figure 4 shows that the estimated posterior distribution was close to the target population distribution, but that it was shifted somewhat to the left of the target population distribution. Figure 5 shows that the estimated posterior distribution was shifted quite a bit to the right of the target population distribution. The means and standard

deviations of the estimated posterior distributions differed from 0 and 1, so the fixing of the item parameters did adjust the scale to some degree. However, that the posterior distributions were not closer to the true population distributions suggests that either insufficient information was available to adequately estimate the group mean and standard deviation, a N(0,1) distribution was still assumed in some computation (i.e., in the likelihood or the prior), or that some kind of rescaling was done on the estimated item parameters. Multilog also has the capability fix certain item parameters and estimate others. Results using the fix capability in Multilog with data generated from a N(-1,1) distribution were virtually identical to the results from Bilog-MG, which suggests that Multilog uses a similar process when item parameters are fixed.

**MML Method for Transforming Item Parameters Appropriate for Sparse CAT Matrices**

Because the bias for the non-fixed items displayed in Figures 2 and 3 appears to be systematic, the parameters for the non-fixed items can be transformed to the scale of the fixed items using a linear transformation. In the conventional data case that we are simulating, we may treat it as if we have a single group taking two mutually exclusive sets of items that are on different scales. If we knew the means and standard deviations of abilities in the group for both scales, we could find the transformation constants to place the non-fixed items on the scale of the fixed items. To do a similar transforming in our CAT data case, we have the added complications of randomly equivalent groups of examinees taking mutually exclusive sets of items, items administered at different frequencies across examinees, and not all examinees taking all items in a pool. We want to place all of our parameters on the scale of one set of items, but there are no common items across the pools, so we cannot use transformation methods that operate on the parameters for common items. Thus, in our case, traditional transformation methods such as the mean/sigma, mean/mean, or characteristic curve methods are not appropriate for obtaining transformation constants for rescaling purposes.

This paper utilizes a method for computing marginal maximum likelihood estimates of transformation constants that place item parameters from one source onto the scale of item parameters from another source. The maximization is conducted with respect to the transformation constants and the population mean and variance. The method can be applied to a single group design, where some common items are shared across examinees (i.e., one group of examinees takes items from the same pool). The method can also be applied to separate, randomly equivalent groups where there are no common items across groups, (i.e., examinees are randomly administered items from one of several pools).

Consider the case where a single group of $N$ examinees is administered $n$ items, where each item has its parameters scaled relative to one of two possible metrics:

$$\delta_i = \begin{cases} 1, & \text{if } i \notin \text{Scale } 1 \\ 0, & \text{if } i \notin \text{Scale } 2, \end{cases}$$

for $i = 1, ..., n$. We further assume that ability is normally distributed $\theta \sim N(\theta; \mu, \sigma)$. Then the likelihood function of the scaling and distribution parameters given the observed responses $\mathbf{u}_a = \{u_{a1}, u_{a2}, ..., u_{an}\}$ for $a = 1, ..., N$ is

$$L(A,B,\mu,\sigma^2|U) = \prod_{a=1}^{N} P(u_a|A,B,\mu,\sigma^2) \qquad (1)$$
$$= \prod_{a=1}^{N} \int P(u_a|\theta,A,B)\, f(\theta|\mu,\sigma^2)\, d\theta,$$

where $f(\theta|\mu,\sigma^2)$ denotes a normal density with mean $\mu$ and variance $\sigma^2$, and where

$$P(u_a|\theta,A,B) = \prod_{i=1}^{n} \left[ P(\theta|a_i,b_i,c_i)^{u_{ai}}\, Q(\theta|a_i,b_i,c_i)^{1-u_{ai}} \right]^{\delta_i}$$
$$\times \prod_{i=1}^{n} \left[ P(\theta|a_i/A, Ab_i+B, c_i)^{u_{ai}}\, Q(\theta|a_i/A, Ab_i+B, c_i)^{1-u_{ai}} \right]^{1-\delta_i}$$

The MML estimates of A and B can be obtained by maximizing the joint likelihood of (1) with respect to A, B, $\mu$ and $\sigma^2$, i.e.,

$$\max_{A,B,\mu,\sigma^2} L(A,B,\mu,\sigma^2|U).$$

This can be accomplished by a two step procedure:

1. Holding A and B fixed, maximize (1) with respect to $\mu$ and $\sigma^2$.
2. Holding $\mu$ and $\sigma^2$ fixed, maximize (1) with respect to A and B.
3. Iterate Steps 1 and 2 until A and B change very little from one iteration to the next.

The process is followed as described above when finding the A and B transformation constants that place Scale 2 parameters on the Scale 1 metric, for a single group of examinees taking some common items. A modified process can be followed to find the A and B transformation constants in the case of randomly equivalent groups being administered mutually exclusive sets of items. In this case, A is fixed to 1.0 and B is fixed to 0.0, and only $\mu$ and $\sigma^2$ are estimated using Equation (1). If $\mu$ and $\sigma^2$ are estimated for the two groups in that manner, the transformation constants to place the Scale 2 parameters on the Scale 1 metric can be obtained by

$$A = \sigma(\text{Scale 1})/\sigma(\text{Scale 2}) \qquad (2)$$
$$B = \mu(\text{Scale 1}) - A*\mu(\text{Scale 2}). \qquad (3)$$

Once the transformation constants A and B have been obtained, the parameters for Scale 2 items can be placed on the metric of Scale 1 items by the following transformations:

$$a^t(\text{Scale 2}) = a(\text{Scale 2})/A \qquad (4)$$
$$b^t(\text{Scale 2}) = A*b(\text{Scale 2}) + B. \qquad (5)$$

**Application of the MML Transformation Method to the Conventional Data**

When the MML transformation method was applied to the conventional data, the following procedure was used. A and B were fixed to 1.0 and 0.0, respectively, and Equation (1) was used to estimate the mean and variance for the group on the 20 "fixed" (or anchor) items, using the

known parameters and the examinee responses to these items. Item parameters for the 74 non-fixed items were obtained by running Bilog-MG on those items only. No attempt was made to fix the scale in this calibration, so results were approximately on the N(0,1) metric. The mean and variance were then estimated for the group on the 74 "non-fixed" items, using the estimated parameters and the examinee responses to these items (and fixing A and B to 1.0 and 0.0, respectively). The transformation constants A and B were obtained using Equations 2 and 3, and the item parameters for the 74 non-fixed items were rescaled applying those transformation constants to Equations 4 and 5.

Figure 6 shows the recovery of the b parameters for the 74 non-fixed items when the MML transformation method was used to rescale parameters for the data simulated from a N(+1,1) distribution. Figure 6 should be compared to Figure 2, which shows the result for the same 74 items when the "fix" capability is used in Bilog-MG to fix the scale of the 74 items to the 20 anchor items. Note that Figure 6 does not represent a transformation of the 74 parameters in Figure 2. Instead, it represents a transformation of the 74 parameters when they were calibrated assuming a N(0,1) scale. The bias that occurred in Figure 2 does not appear in Figure 6. Figure 7 shows the recovery of the b parameters for the non-fixed items when the MML transformation method was used to rescale parameters for the data simulated from a N(-1,1) distribution. Figure 7 should be compared to Figure 3, which shows the results for the same items using the "fix" capability in Bilog-MG. Again, the bias that was evident in Figure 3 does not appear in Figure 7.

Figure 8 and 9 show the root mean squared difference (RMSD) between the item characteristic curves (ICCs) based on the estimated and true parameters, for the data simulated from a N(+1,1) and N(-1,1) distribution, respectively. For each item, the estimated and true ICCs were computed at the true ability for each examinee taking that item, and the sum of the squared difference in ICCs was weighted by the number of examinees taking that item prior to taking the square root. The results labeled "Fixed" are based on the estimated parameters for the 74 items when the "fix" capability is used in Bilog-MG. The results labeled "Rescaled" are based on the rescaled parameters for the 74 items when the MML transformation method is applied to the N(0,1) scaled Bilog-MG parameters. Figure 8 and 9 demonstrate a substantial decrease in RMSD from the fixed parameters to the rescaled parameters. For comparison purposes, Figure 10 shows the RMSD results for the data simulated from a N(0,1) distribution. In this case, the Fixed solution and the Rescaled solution give very similar results. Table 2 summarizes the average RMSD over the 74 items for the Fixed and Rescaled parameters.

Table 2. Average RMSD for the Fixed and Rescaled Parameters, by Examinee Ability Distribution.

| Examinee Distribution | Fixed | Rescaled |
|---|---|---|
| N(0,1) | .007 | .007 |
| N(1,1) | .024 | .010 |
| N(-1,1) | .040 | .013 |

In total, the results for the conventional data all suggest that using the Bilog-MG fix capability to place estimated parameters on the scale of the fixed items does not place them quite on the

desired scale, when the examinees do not conform to a N(0,1) ability distribution. The RMSDs for the N(0,1) data show us the optimal RMSD we can expect using the Bilog-MG fix capability, when the underlying assumptions made by the program are met in the data. This is the target RMSD we would like to achieve. Unfortunately, when the examinees are not distributed N(0,1), the results from using the Bilog-MG fix capability are biased to some degree.

In this example, the percentage of fixed items used was comparable to the percentage of fixed items that we would have for our operational CAT. Increasing the number of fixed items when using the fix capability does reduce the bias in the parameter estimates, but a very large number of items need to be fixed before the RMSDs are reduced to a comparable level as observed for the N(0,1) parameters that are rescaled using the MML transformation method. Figure 11 shows the RMSD over 44 common items when 20, 30, 40, and 50 items are fixed in Bilog-MG, for the data simulated from a N(-1,1) distribution. Table 3 summarizes the average RMSD over the 44 common items when using the Bilog-MG fix capability with different numbers of fixed items (labeled Fixed). The average RMSD over the same 44 items for the N(0,1) scaled parameters that are rescaled to the scale of the 20 fixed items is presented for comparison (labeled Rescaled). When over half of the items are fixed (50 out of 94), the average RMSD for the non-fixed parameters from the Bilog-MG "fixed" solution begins to approach the value of the average RMSD observed when the MML transformation method was used to rescale the item parameters to the scale of 20 anchor items.

Table 3. Average RMSD over 44 Common Items by Number of Fixed Items.

| Number Fixed Items | Fixed | Rescaled |
|---|---|---|
| 20 | .037 | .012 |
| 30 | .025 | - |
| 40 | .019 | - |
| 50 | .016 | - |

**Evaluation of CAT 3PL Data**

The CAT data were simulated as discussed earlier. Because of the randomly equivalent groups, the calibration can be conducted as a single group analysis on one set of items. Thus, the calibration problem consisted of 122400 examinees and 605 items (with each examinee taking 16 items). Items that were not administered to a given examinee were treated as not presented and ignored in the calibrations. This created a large sparse matrix for analysis. For the data generated from a N(0,1) distribution, item administration rates for Pools 1-3 ranged from 5 administrations of one item to 14,283 administrations of another. Administration rates for Pool 4, which was administered only about 2% of the time, ranged from 0 administrations of one item to 825 administrations of another. Because the parameters for Pool 4 are never re-estimated, the small sample size associated with this pool was not a concern. Items were administered at different rates across the N(0,1), N(+1,0.8), and N(-1,1.2) conditions.

Figure 12 shows the recovery of the b parameters for the non-fixed pretest items when Bilog-MG was used to fix the parameters for the anchor items (Pool 4) and estimate the parameters for the

remainder of the items (Pools 1-3 and the pretest items), and the data were generated from a N(0,1) distribution. In this case, there were 137 fixed items and 468 non-fixed items. The results are plotted only for the 100 pretest items. Although we would like the calibration of the Pool 1-3 items to be the best possible, we are ultimately only concerned with the quality of the calibrations for the pretest items, as the parameters for the operational pools will not be updated during the life of a pool. Figure 12 shows that the parameter recovery is not as good for the CAT case as in the case of the conventional N(0,1) data (see Figure 1). The smaller sample size for these items (N=1224 for the CAT case versus N=10,000 for the conventional case) is one likely cause of this. However, another possible contributor could be the poor calibration of some of the Pool 1-3 items.

Some Pool 1-3 items were calibrated poorly simply because very few examinees took them. Others were calibrated poorly because the starting values for the parameters were too far from the true parameters for Bilog-MG to converge on the appropriate solution. Bilog-MG uses classical item statistics (biserial correlation and percent correct) as default starting values for the a and b parameters. The classical item statistics provide good starting values for conventional data, but not for adaptive CAT data. Biserial correlations are often negative in the case of CAT data, and percent corrects may be based on a subset of examinees that are at similar levels of ability. In some cases, it is necessary to over-ride the default starting values in order to get Bilog-MG to even run with sparse CAT data. In this study, starting values of 1.0 were used for the a parameter and starting values of 0.0 were used for the b parameter, for all items in Pools 1-3. For some of these items, Bilog-MG could not recover from start values for the b parameters that were too far from the true parameters. Thus, some individual operational items were estimated very poorly.

Assigning appropriate starting values for a sparse CAT data matrix is an interesting problem that was not explicitly addressed in this study. Fortunately, only a small percentage of Pool 1-3 items were affected by poor start values. If we were concerned with obtaining the optimal calibration possible for Pools 1-3 items, we would have to find a better way of assigning start values to those items. Since our concern was with the optimal calibration of the pretest items, we were able to overlook the noise that was caused by a few poorly calibrated items from Pools 1-3. Fortunately, starting values were only a concern for the items that were administered adaptively. The default start values used in Bilog-MG were appropriate for the pretest items because those items were administered at random to examinees and were not targeted toward a particular ability.

Bilog-MG uses two methods of estimating the parameters: the EM method is implemented first, and upon convergence, is followed by the Newton-Gauss (Fisher scoring) method. For the CAT data, the Newton-Gauss steps were suppressed because the results did not converge for any of the data. The results all converged for the EM steps, unless otherwise reported. Because of the nature of the data, it was also necessary to use a prior distribution when calibrating the b parameters for the sparse CAT data, in order to obtain a converged solution. The default prior for the b parameters was used. The calibration problems noted here only occurred with the CAT data. For the conventional data, there were no convergence problems.

Figures 13-14 show the recovery of the b parameters for the non-fixed pretest items when Bilog-MG was used to fix the parameters for the anchor items (Pool 4) and estimate the parameters for the remainder of the items (Pools 1-3 and the pretest items), and the data were generated from a N(+1,0.8) and N(-1,1.2) distribution, respectively. (Note that two items are excluded from Figure 13, each with an estimated b parameter of < -7.0.) Trends occur in the same direction as those observed for the comparable distribution in the conventional data case (see Figures 2 and 3). Namely, the estimated parameters were underestimated for the data generated from the N(+1,0.8) distribution, and the estimated parameters were inflated for the data generated from the N(-1,1.2) distribution. As might be expected, the magnitude of the bias was much greater for the sparse CAT data than for the conventional data.

Figures 15-16 show the recovery of the b parameters for the 100 non-fixed pretest items when the MML transformation method was used to rescale parameters for the data simulated from a N(+1,0.8) and N(-1,1.2) distribution, respectively. (Note that two items are excluded from Figure 15, each with an estimated b parameter of < -5.0.) When the MML transformation method was applied to the CAT data, the following procedure was used. A and B were fixed to 1.0 and 0.0, respectively, and Equation (1) was used to estimate the mean and variance for the group on the 137 "fixed" (or anchor) items from Pool 4, using the known parameters and the examinee responses to these items. Item parameters for the 468 non-fixed items (Pools 1-3 and the pretest items) were obtained by running Bilog-MG on those items only. No attempt was made to fix the scale in this calibration, so results were approximately on the N(0,1) metric. The mean and variance for the group on the non-fixed items was assumed to be 0.0 and 1.0, respectively. The transformation constants A and B were obtained using Equations 2 and 3, and the item parameters for the 468 non-fixed items were rescaled applying the transformation constants to Equations 4 and 5.

A comparison of Figures 13 and 15 shows that the bias that occurred from using Bilog-MG to fix the scale did not occur when MML transformation method was used to rescale the non-fixed parameters when they were calibrated to a N(0,1) scale. The rescaled parameters do show more spread than in the conventional data case, which is a reflection of the poorer calibration that occurred under the sparse CAT data simulation. A comparison of Figures 14 and 16 also shows that the bias that occurred from using Bilog-MG to fix the scale was reduced when the MML transformation method was used to rescale the non-fixed parameters after they were calibrated to a N(0,1) scale. The rescaled parameters do show a slight amount of bias, but that may have been caused by the quality of the calibration. In the calibration of the non-fixed parameters (i.e., the 468 items) to a N(0,1) scale, Bilog-MG initially approached convergence, but then bounced around and did not reach a convergence criterion of .001 after 100 iterations (in the EM step). The parameters were probably not as well estimated as they would have been if the results had converged. Manipulations of the Bilog-MG options did not result in a converged solution. The parameters reported here are the parameters after 100 iterations. No convergence problems were noted in the N(+1,0.8) case.

Figures 17 and 18 show the root mean squared difference (RMSD) between the item characteristic curves (ICCs) based on the estimated and true parameters for the 100 pretest items, for the CAT data simulated from a N(+1,0.8) and N(-1,1.2) distribution, respectively. For each item, the estimated and true ICCs were computed at the true ability for each examinee taking that item, and the sum of the squared difference in ICCs was weighted by the number of examinees

taking that item prior to taking the square root. The results labeled "Fixed" are based on the estimated parameters for the 100 pretest items when the "fix" capability was used in Bilog-MG (recall parameters were estimated for 468 items, although only results for the 100 pretest items are presented). The results labeled "Rescaled" are based on the rescaled parameters for the 100 items when the MML transformation method was applied to the N(0,1) scaled Bilog-MG parameters for the 468 items. Figures 17 and 18 demonstrate a substantial decrease in RMSD from the fixed parameters to the rescaled parameters.

For comparison purposes, Figure 19 shows the RMSD results for the CAT data simulated from a N(0,1) distribution. As was observed with the conventional data from a N(0,1) distribution (see Figure 10), the Fixed solution and the Rescaled solution give very similar results. Table 4 summarizes the average RMSD over the 100 items for the Fixed and Rescaled parameters. The average RMSD was larger for the rescaled parameters in the N(-1,1.2) data, but this is likely because the parameters were as not well estimated as they could have been (recall that the results did not converge in the calibration).

Table 4. Average RMSD for 100 Pretest Items, for the Fixed and Rescaled Parameters, by Examinee Ability Distribution.

| Examinee Distribution | Fixed | Rescaled |
|---|---|---|
| N(0,1) | .018 | .018 |
| N(+1,0.8) | .159 | .018 |
| N(-1,1.2) | .157 | .036 |

## Discussion

This paper evaluated two different methods of calibrating non-anchor items to the scale of anchor items. The first method used Bilog-MG to fix parameters for anchor items to their known values and simultaneously estimate the parameters for non-anchor items. The second method used Bilog-MG to estimate parameters for non-anchor items only (without attempting to fix the scale), and then used a MML transformation procedure to rescale those parameters to the scale of the anchor items. The results suggest that it is safer to calibrate the non-anchor items to a N(0,1) scale and then rescale to the scale of the anchor items, rather than to fix the parameters for the anchor items and simultaneously calibrate the non-anchor items. If the examinees are distributed N(0,1), the two procedures will give similar results, but it could be difficult in practice to know for certain how the examinees are distributed.

If the examinee ability distribution is shifted in location from N(0,1), fixing the parameters for the anchor items and calibrating the non-anchor items could result in biased estimates. Results from using the "fix" capability in Bilog-MG were biased for both the conventional and CAT data, when the examinee ability distributions were shifted away from N(0,1). The bias in the estimates for the non-anchor items resulting from fixing the anchor item parameters was much larger for the CAT data than for the conventional data because of the small number of examinees taking the anchor items relative to the rest of the items to be estimated. There was very little information for Bilog-MG to use in fixing the scale. The MML transformation method worked

well even in the case of the CAT data, because it did not re-estimate any item parameters. Thus, the smaller sample sizes for the anchor items were not a concern.

All attempts to fix the scale of the non-anchor items to the scale of the anchor items using Bilog-MG options were unsuccessful. A variety of different options were tried, but none enabled the recovery of the original scale. Ideally, if some item parameters are fixed, the population mean and standard deviation would be estimated using the fixed parameters and the responses to those items. That estimated prior distribution would then be used in every computation that uses a prior, rather than using a N(0,1) prior. The bias observed when using the "fix" capability in Bilog-MG suggests that somewhere a N(0,1) distribution is being assumed, or that not enough information is available to adequately fix the scale. There appears to be no obvious way when using the fix capability in Bilog-MG to ensure that the non-anchor items are truly on the scale of the anchor items.

Parameter recovery for the pretest items was promising when the MML transformation method was used to rescale parameters for the non-anchor items onto the scale of the anchor items. The parameter recovery for the operational pools (Pools 1-3) was not presented in this paper. However, some of the operational items were calibrated very poorly. Our calibration problem presented somewhat of an optimal calibration situation, as we were not overly concerned with the quality of the calibration of the operational CAT items, even though those items were simultaneously calibrated with the pretest items. Clearly we would like the operational items to be calibrated as best as possible, but an occasional poor calibration of some operational items did not appear to have much of an effect on the calibration of the pretest items. If our interest was also in the calibration of the operational items, then the calibration problem would be much more difficult.

There are some calibration issues that arise when dealing with sparse CAT data. Small sample sizes for individual items can be problematic. The Newton-Gauss steps do not appear to work with the CAT data. However, identifying appropriate starting values is probably the critical issue. The default starting values used in Bilog-MG were suitable for our pretest items because they were administered randomly across examinees. When default start values were used for the operational items, Bilog-MG had problems running. When start values of a = 1.0 and b = 0.0 were used instead of the defaults for the operational items, Bilog-MG was able to run. In some instances, however, the starting values for the b parameters were too far from the true parameters, and Bilog-MG converged on an incorrect solution. It was easy to detect this occurrence with simulated data, but would be difficult to do so with real data. The diagnostic $\chi^2$ test given for each item was generally very large for the operational CAT items, so it would be difficult to use that test to detect that an inappropriate starting value had been used.

Although results were promising for the parameters that were rescaled using the MML transformation method, the results for the rescaled parameters will be only as good as the underlying unscaled parameters. In the case of the N(-1,1.2) CAT data, the parameters did not converge, and the RMSDs for the rescaled pretest parameters were larger than observed for the N(+1,0.8) CAT data. If it is not possible to obtain a converged solution when calibrating, then it might not be appropriate to apply the MML transformation method to those results. Calibrations of CAT data may be affected not only by the sparseness of the data, but by the relationship

between the item and the ability of the calibration sample. For example, in the case of a very difficult item that is administered to a below-average ability group, it may be difficult to calibrate that item, no matter how many people take it.

The simulations presented here were designed to demonstrate parameter recovery when items all fit a 3PL model. Because Bilog-MG and the MML transformation method both assume a 3PL model, our results represent the optimal evaluation of these methods. In cases where the 3PL model is not appropriate, the results are likely to show more error. The final stage of the simulation study will be to evaluate the methods when some items do not fit a 3PL model.

## References


du Toit, M. (2003). *IRT from SSI*. Lincolnwood, IL: Scientific Software International, Inc.

Krass, I.A., & Williams, B. (2003). *Calibrating CAT Pools and Online Pretest Items Using Nonparametric and Adjusted Marginal Maximum Likelihood Methods.* Paper presented at the Annual Meeting of the National Council on Measurement in Education, Chicago, IL.

Muraki, E., & Bock. R.D. (2003). Parscale. In M. du Toit (Ed.), *IRT from SSI* (pp. 257-344). Lincolnwood, IL: Scientific Software International, Inc.

Nicewander, A. (2003). *Issues in maintaining scale consistency for the CAT-ASVAB.* Paper presented at the Annual Meeting of the National Council on Measurement in Education, Chicago, IL.

Segall, D.O. (2003). *Calibrating CAT pools and online pretest items using MCMC methods*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Chicago, IL.

Thissen, D. (2003). Multilog. In M. du Toit (Ed.), *IRT from SSI* (pp. 345-409). Lincolnwood, IL: Scientific Software International, Inc.

Thomasson, G. (2003). Evaluating the stability of online item calibrations under varying conditions. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Chicago, IL.

Zimowski, M., Muraki, E., Mislevy, R., & Bock, R.D. (2003). In M. du Toit (Ed.), *IRT from SSI* (pp. 24-256). Lincolnwood, IL: Scientific Software International, Inc.

Figure 1. Recovery of b Parameters for a Bilog-MG Calibration Fixing 20 Parameters and Estimating 74 Parameters, for Conventional Data Simulated from a N(0,1) Distribution.

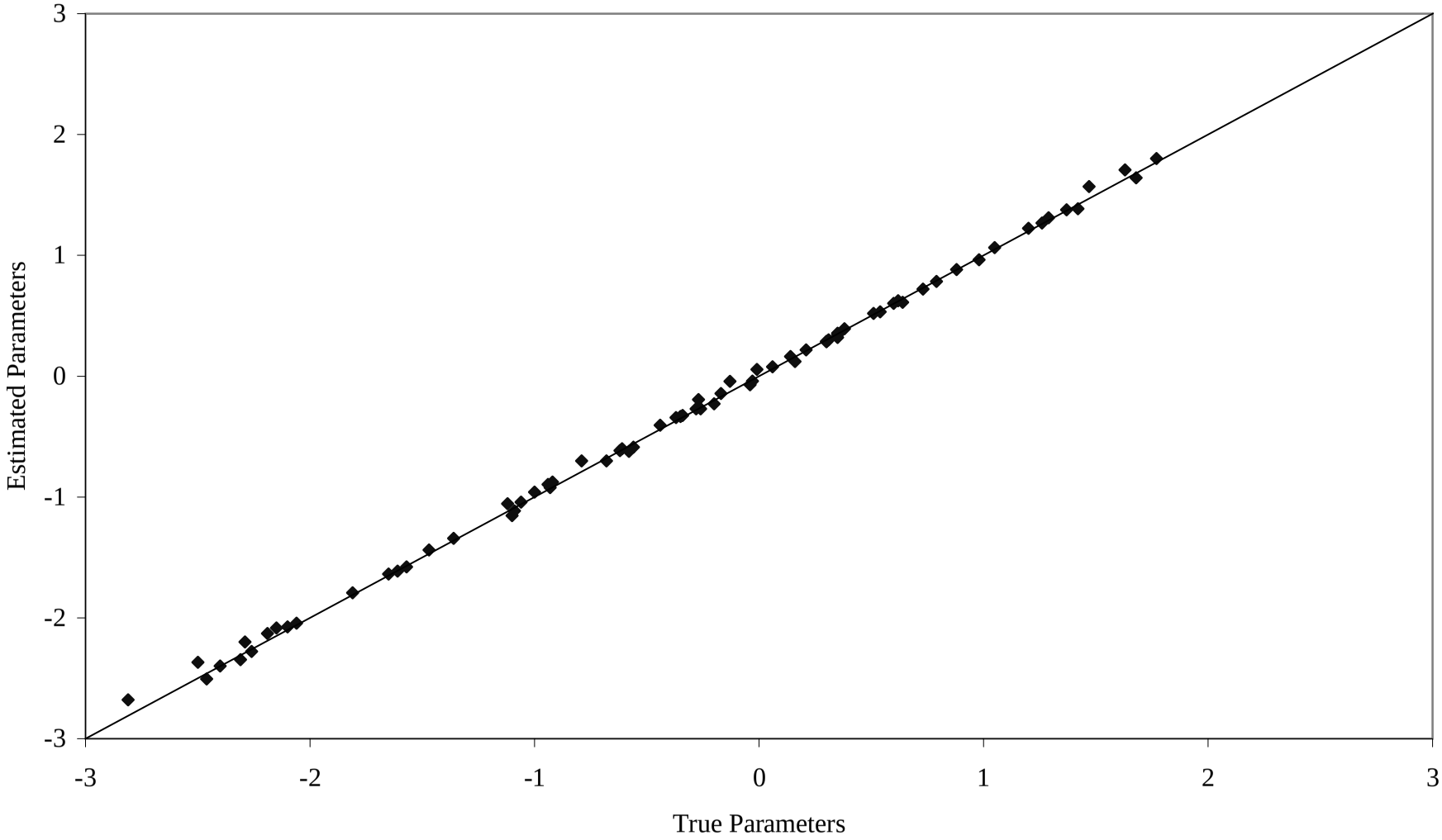


Figure 2. Recovery of b Parameters for a Bilog-MG Calibration Fixing 20 Parameters and Estimating 74 Parameters, for Conventional Data Simulated from a N(+1,1) Distribution.

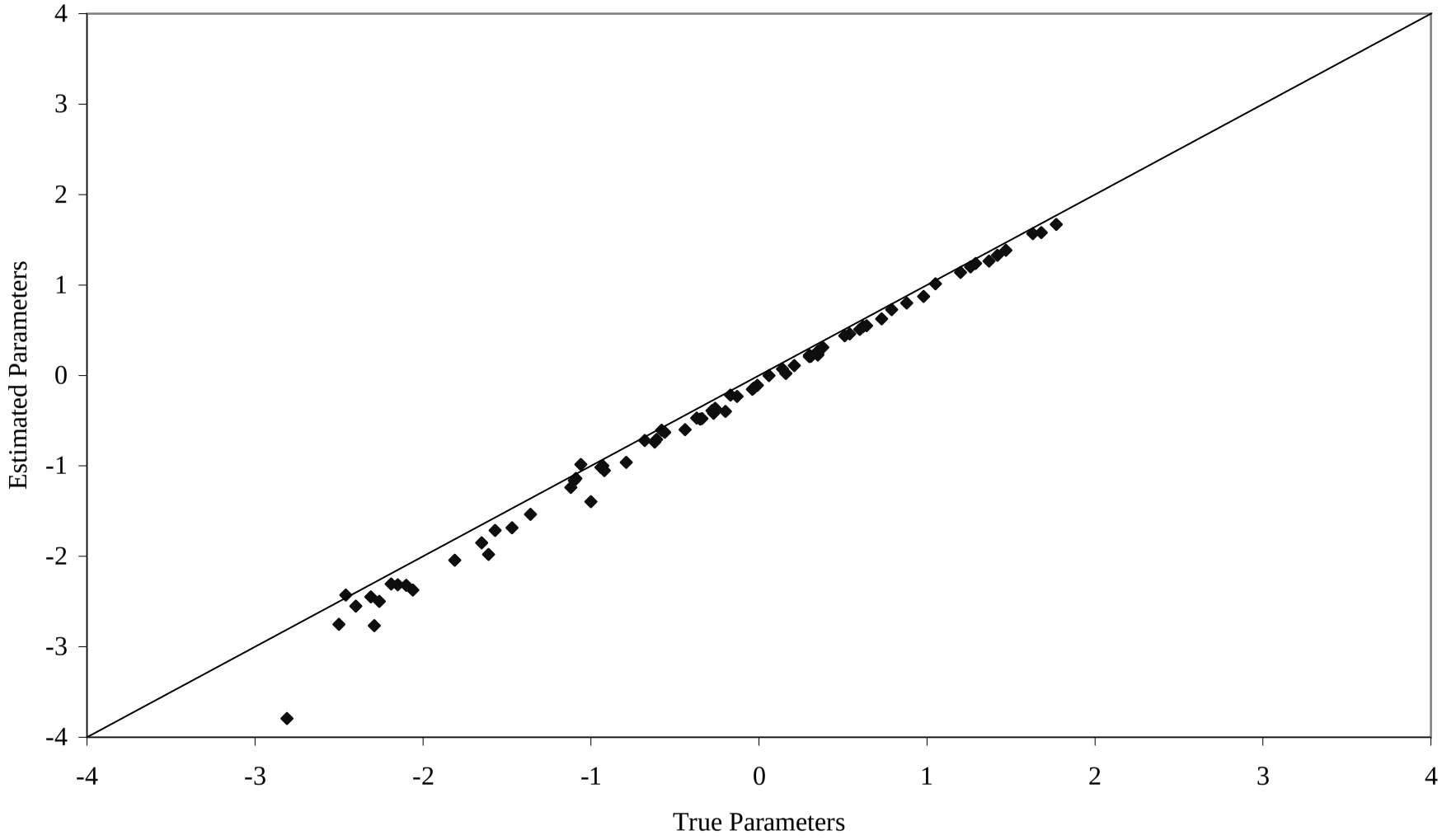


Figure 3. Recovery of b Parameters for a Bilog-MG Calibration Fixing 20 Parameters and Estimating 74 Parameters, for Conventional Data Simulated from a N(-1,1) Distribution.

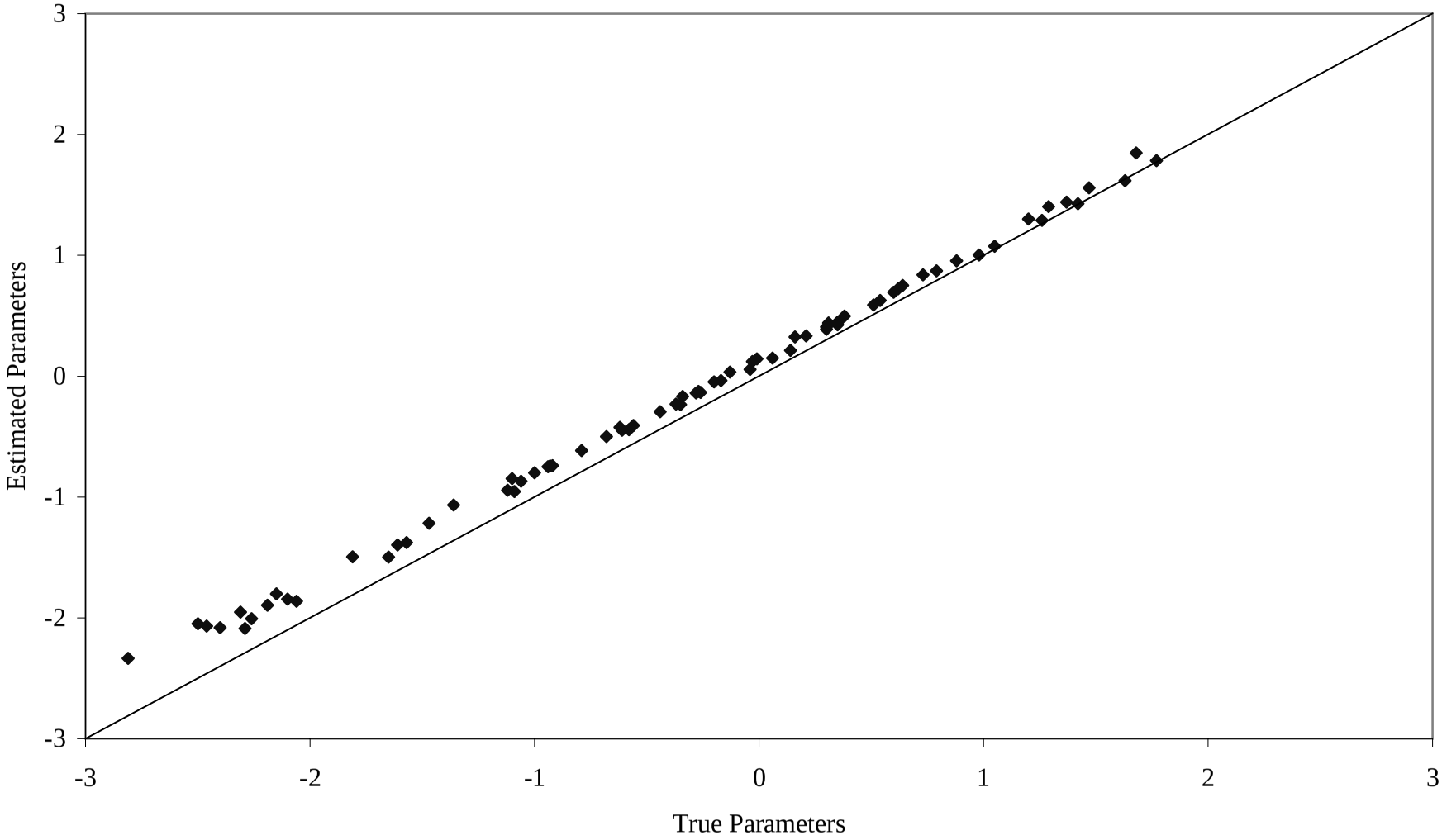


Figure 4. Estimated Posterior Distribution from a Bilog Calibration Fixing 20 Parameters and Estimating 74 Parameters, for Conventional Data Simulated from a N(+1,1) Distribution.

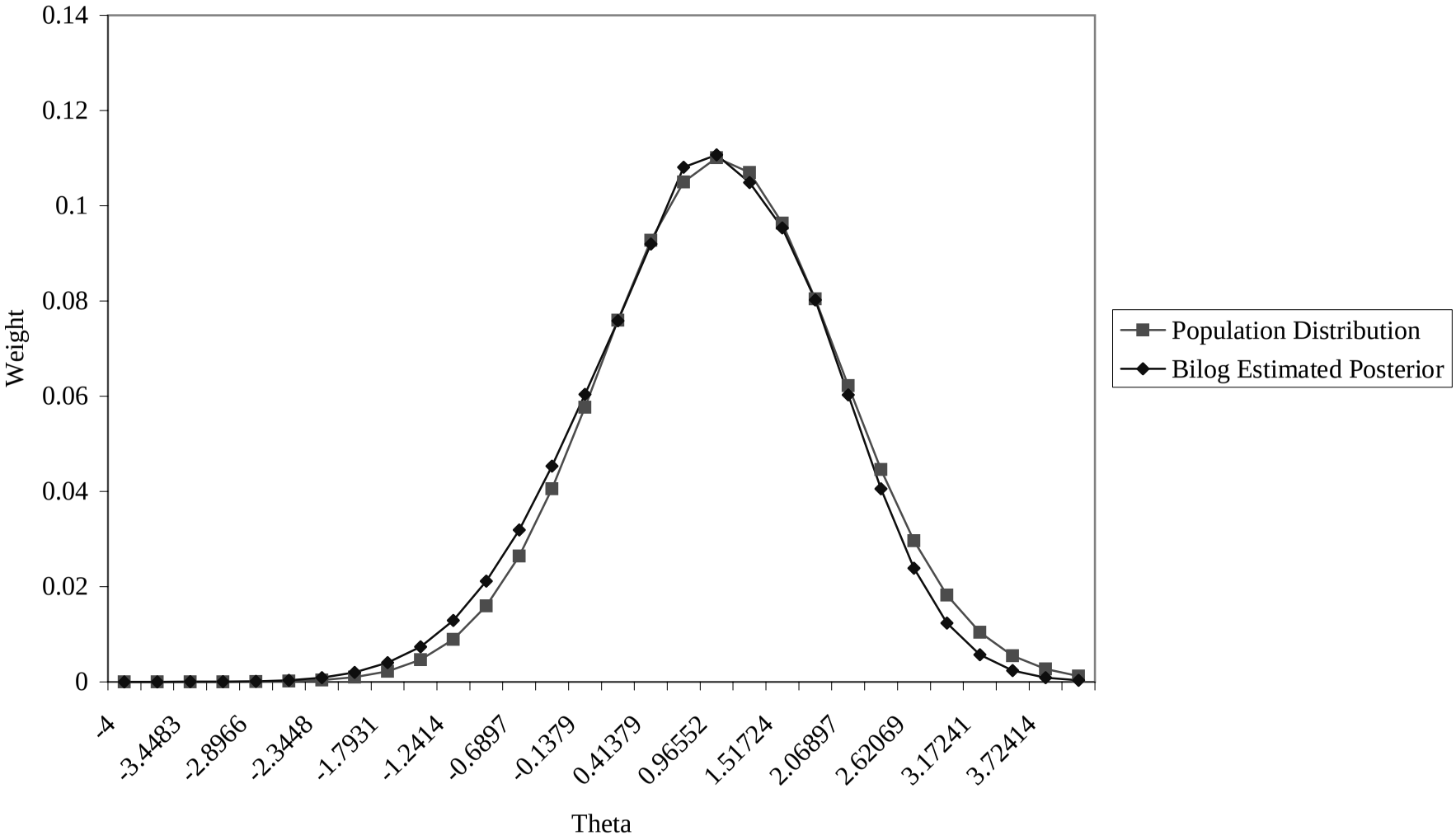


Figure 5. Estimated Posterior Distribution from a Bilog Calibration Fixing 20 Parameters and Estimating 74 Parameters, for Conventional Data Simulated from a N(-1,1) Distribution.

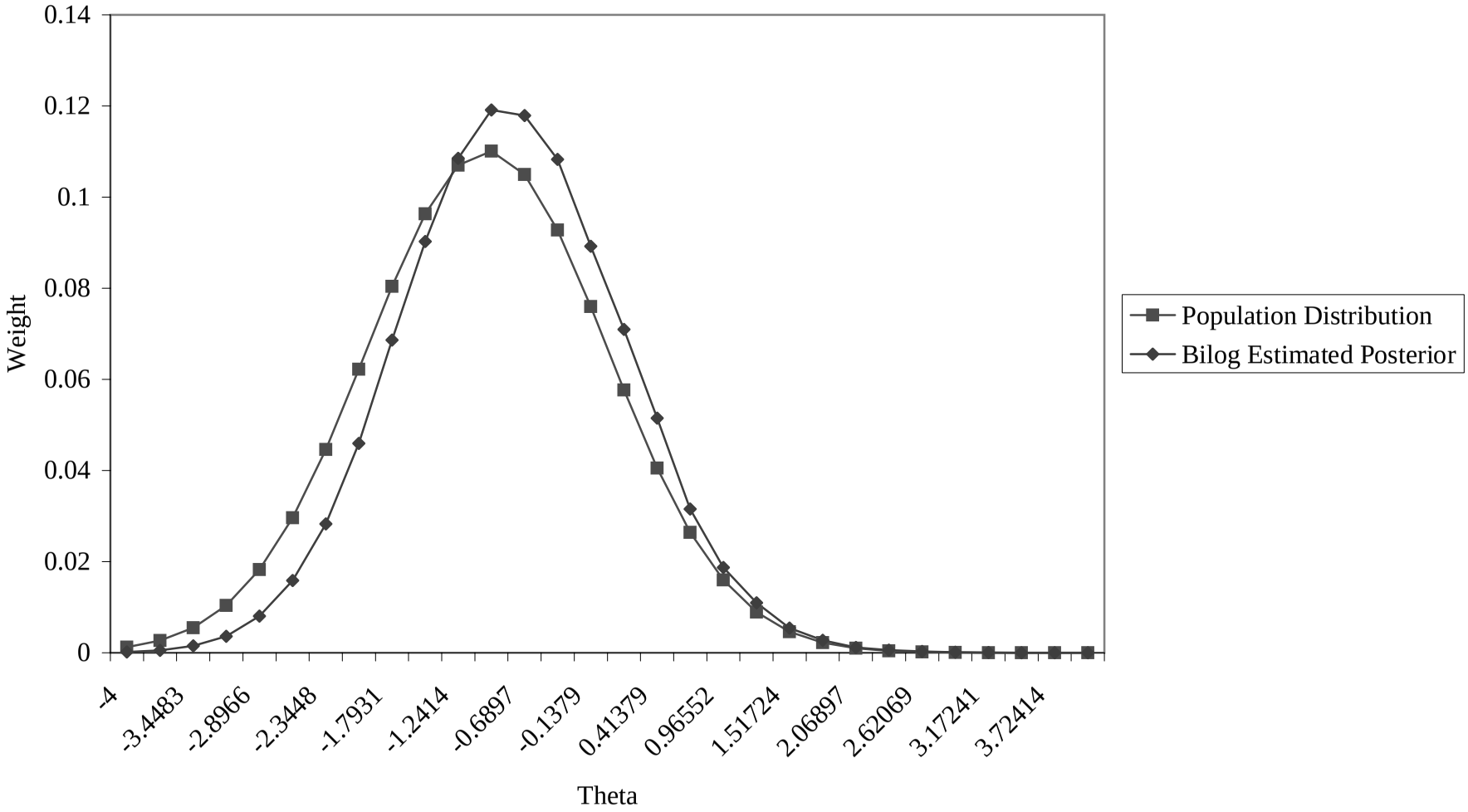


Figure 6. Recovery of b Parameters when the MML Transformation Procedure was Used to Rescale the Parameters, for Conventional Data Simulated from a N(+1,1) Distribution.

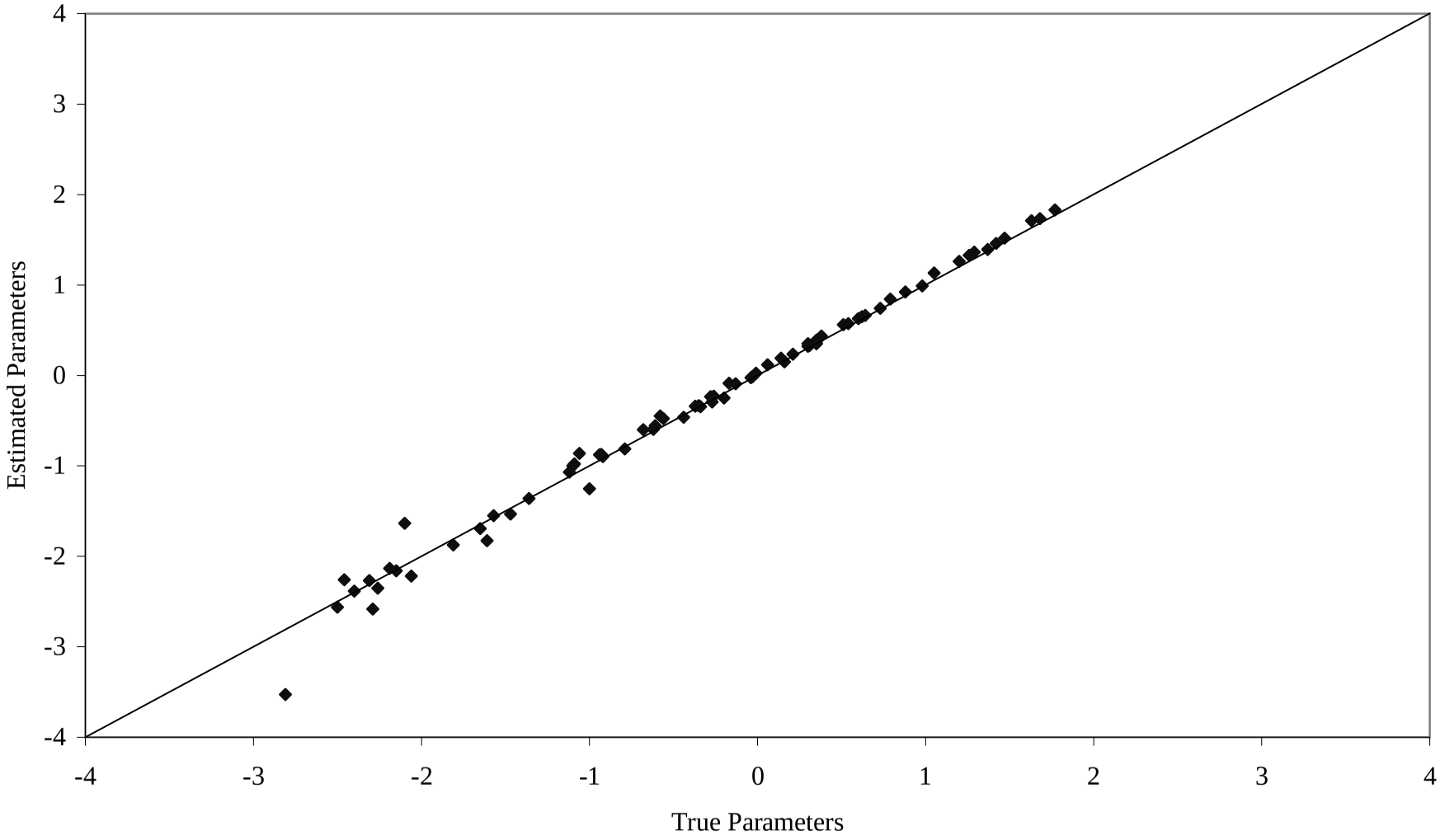


Figure 7. Recovery of b Parameters when the MML Transformation Procedure was Used to Rescale the Parameters, for Conventional Data Simulated from a N(-1,1) Distribution.

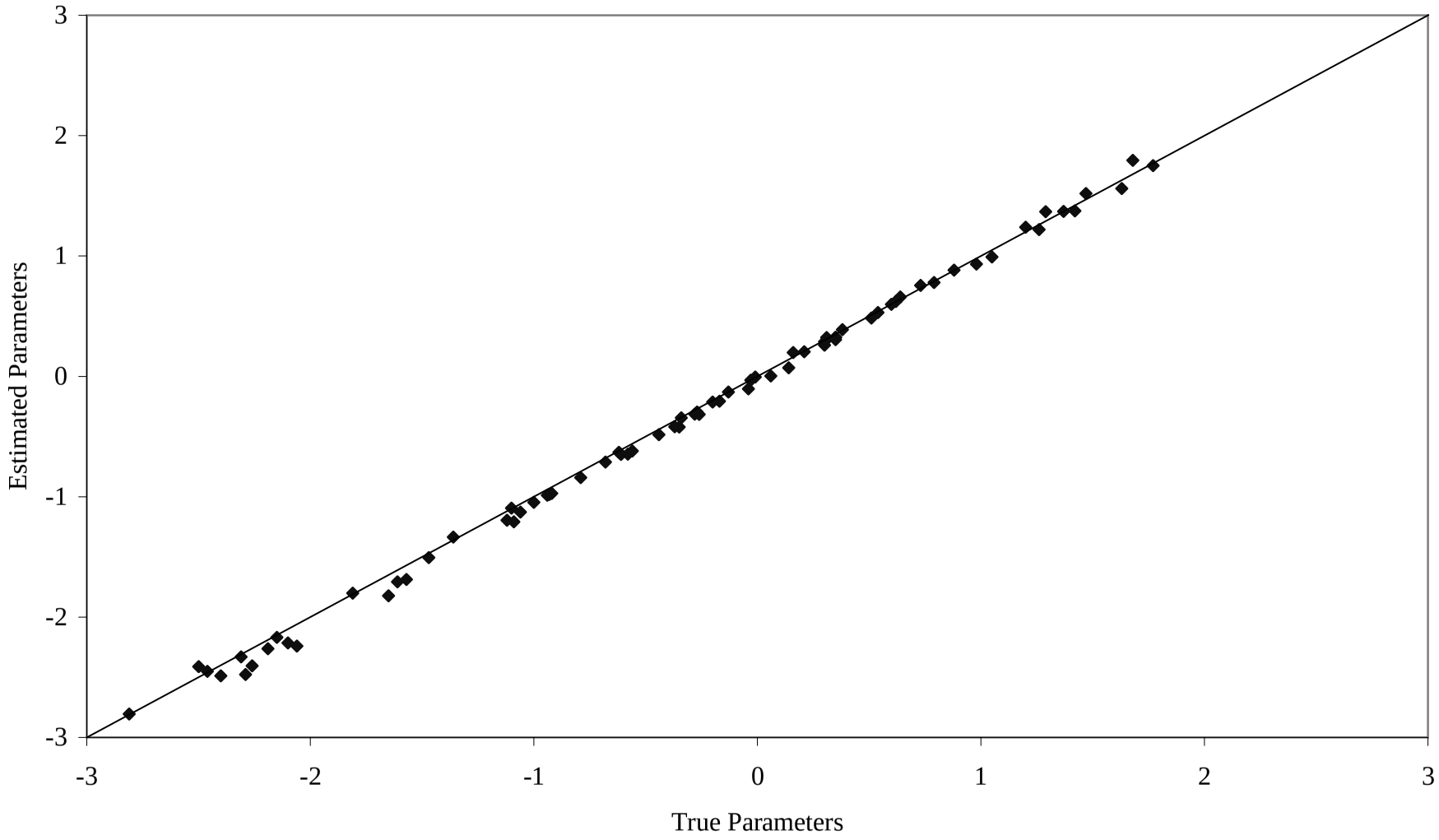


Figure 8. RMSD for Parameters Estimated Using the Bilog-MG Fix Capability and Parameters that are Rescaled Using the MML Transformation Procedure, for Conventional Data Simulated from a N(+1,1) Distribution.

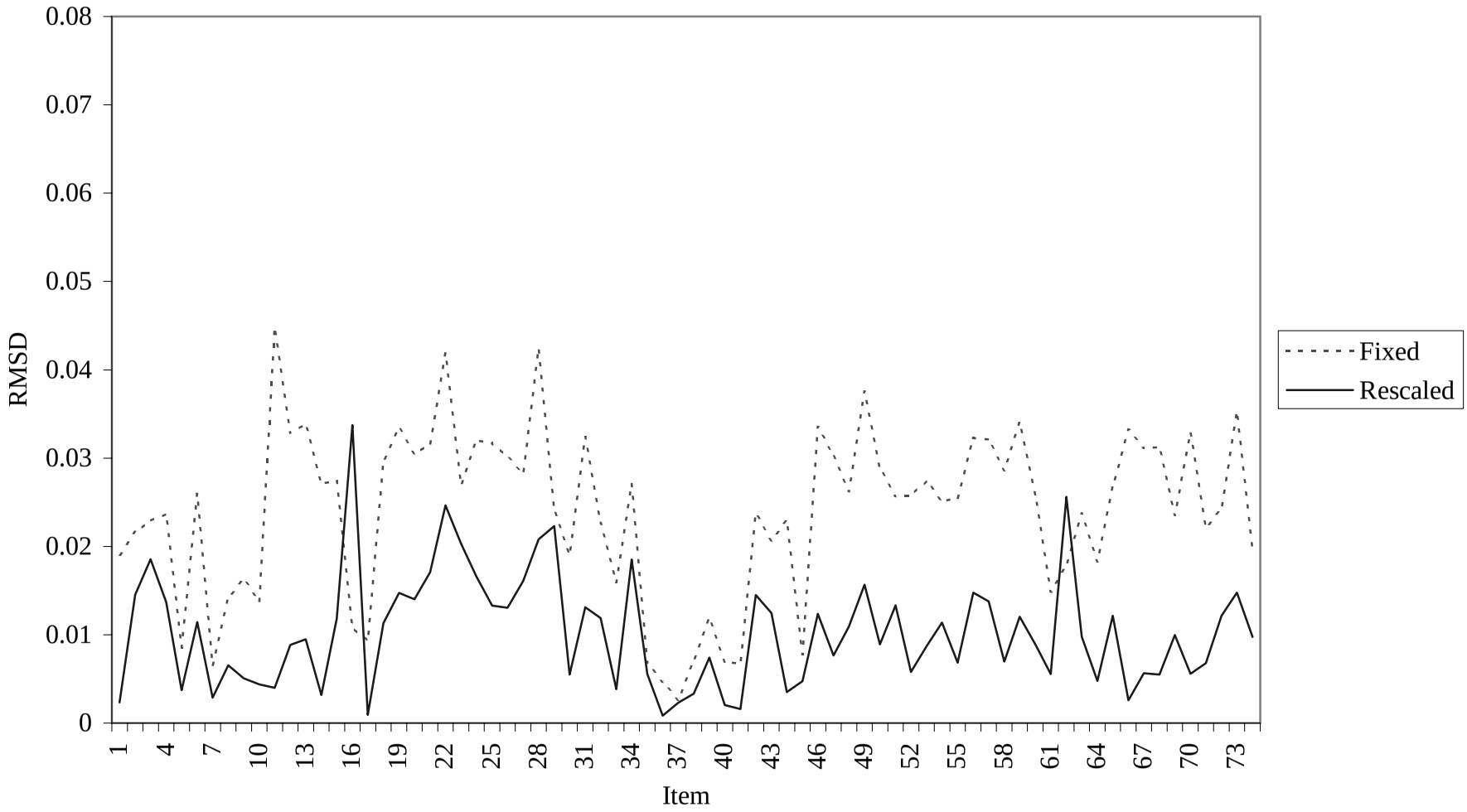


Figure 9. RMSD for Parameters Estimated Using the Bilog-MG Fix Capability and Parameters that are Rescaled Using the MML Transformation Procedure, for Conventional Data Simulated from a N(-1,1) Distribution.

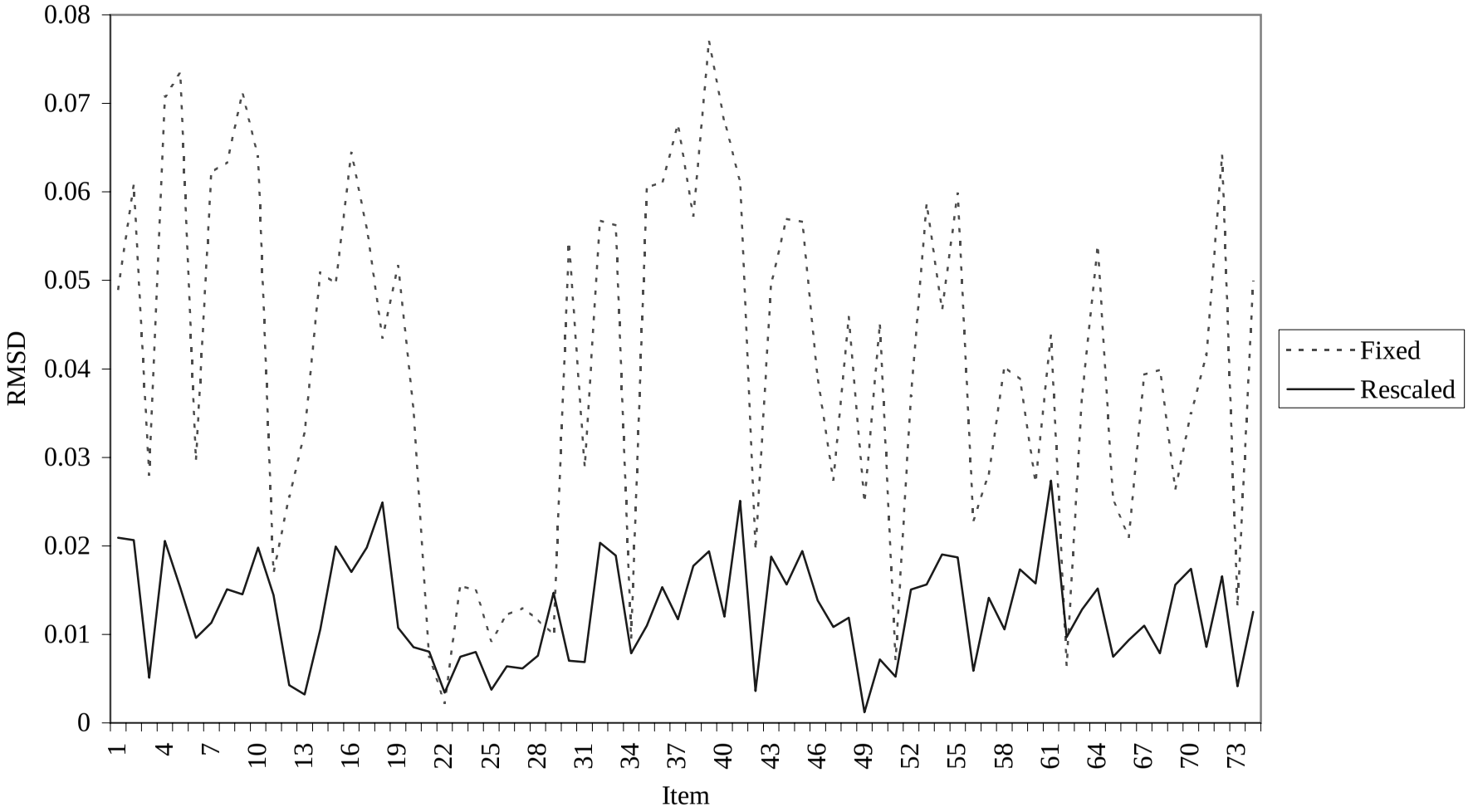


Figure 10. RMSD for Parameters Estimated Using the Bilog-MG Fix Capability and Parameters that are Rescaled Using the MML Transformation Procedure, for Conventional Data Simulated from a N(0,1) Distribution.

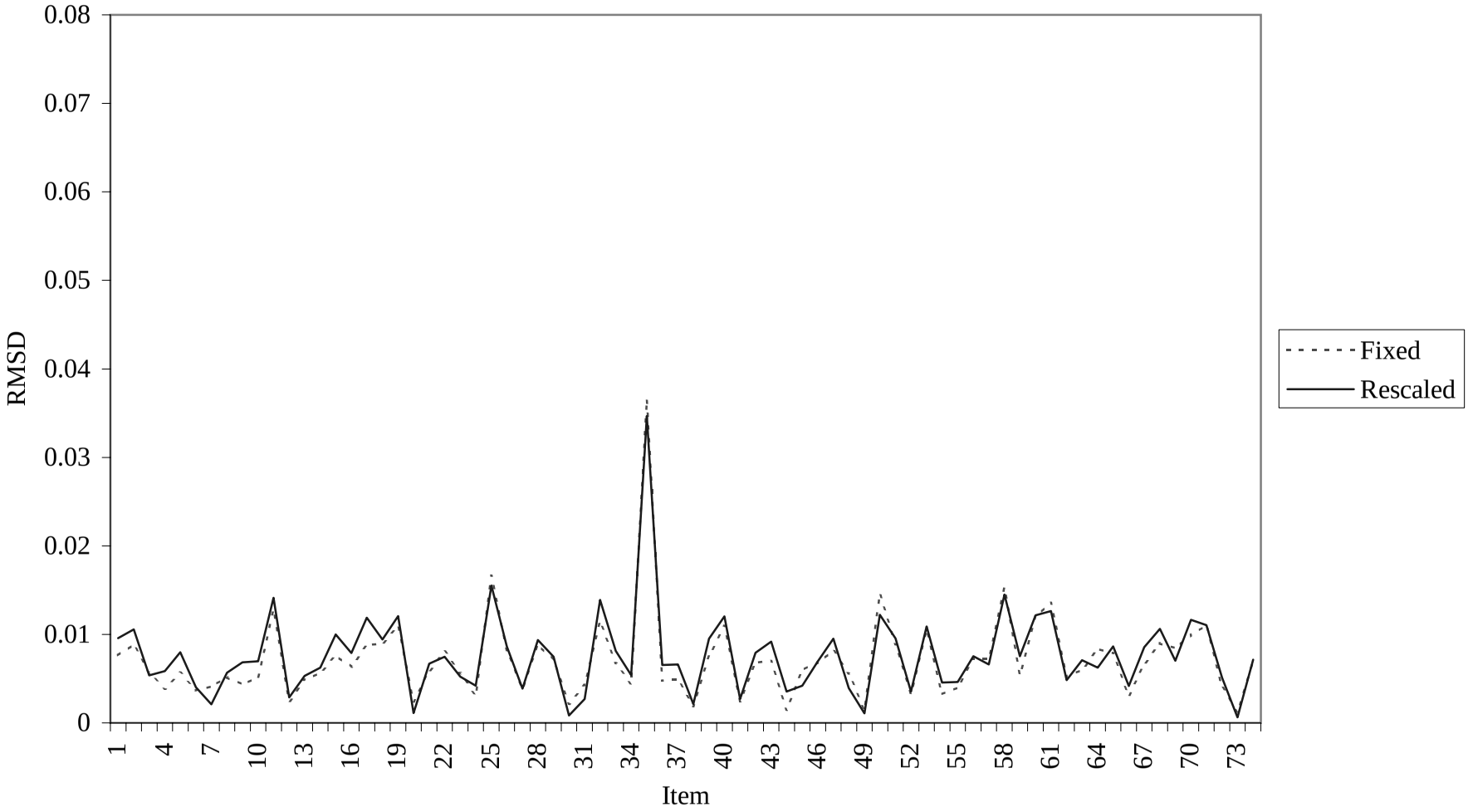


Figure 11. RMSD using the Bilog-MG Fix Capability with Different Numbers of Fixed Items, for Conventional Data Simulated from a N(-1,1) Distribution.

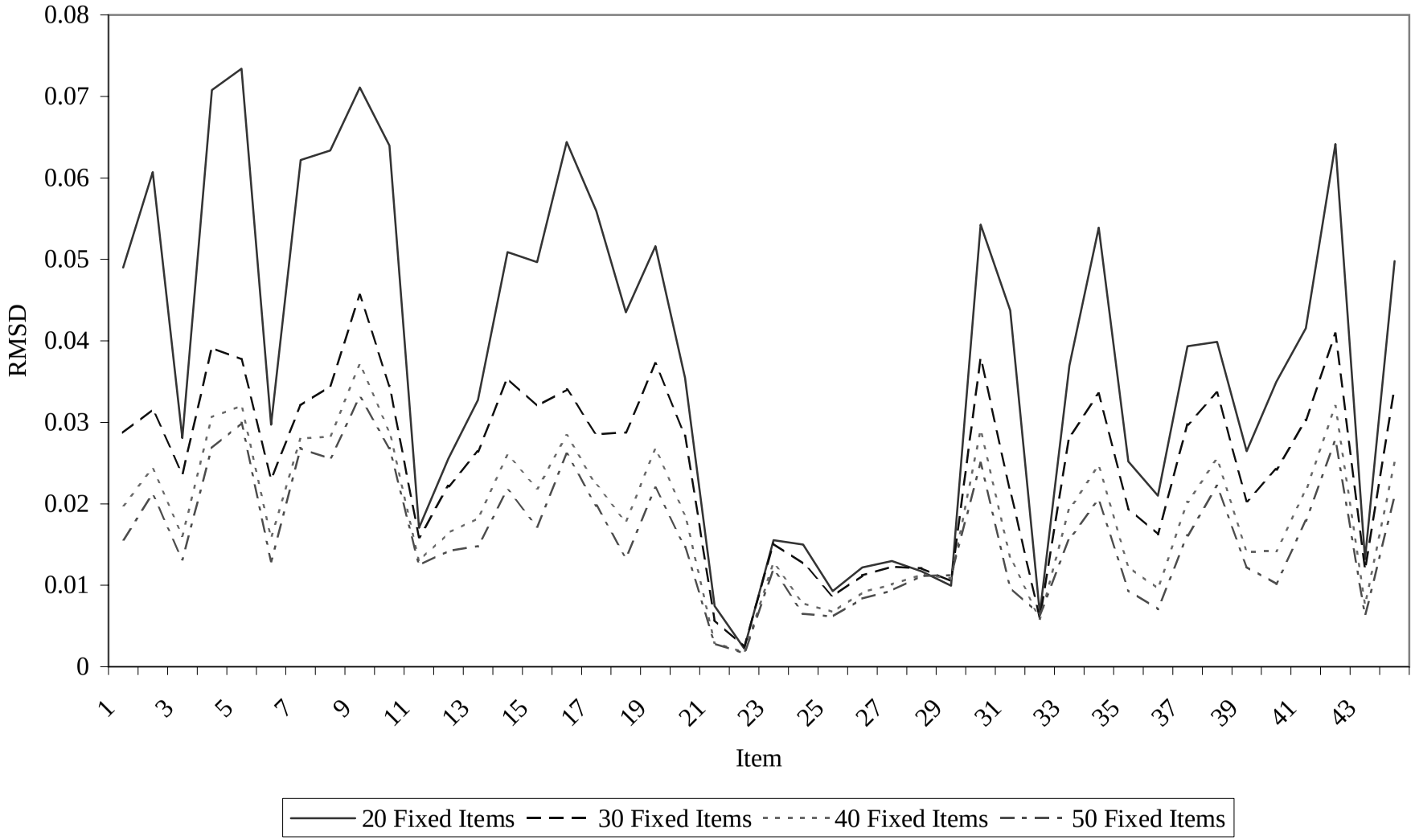


Figure 12. Recovery of b Parameters for 100 Pretest Items from a Bilog-MG Calibration Fixing 137 Parameters and Estimating 468 Parameters, for CAT Data Simulated from a N(0,1) Distribution.

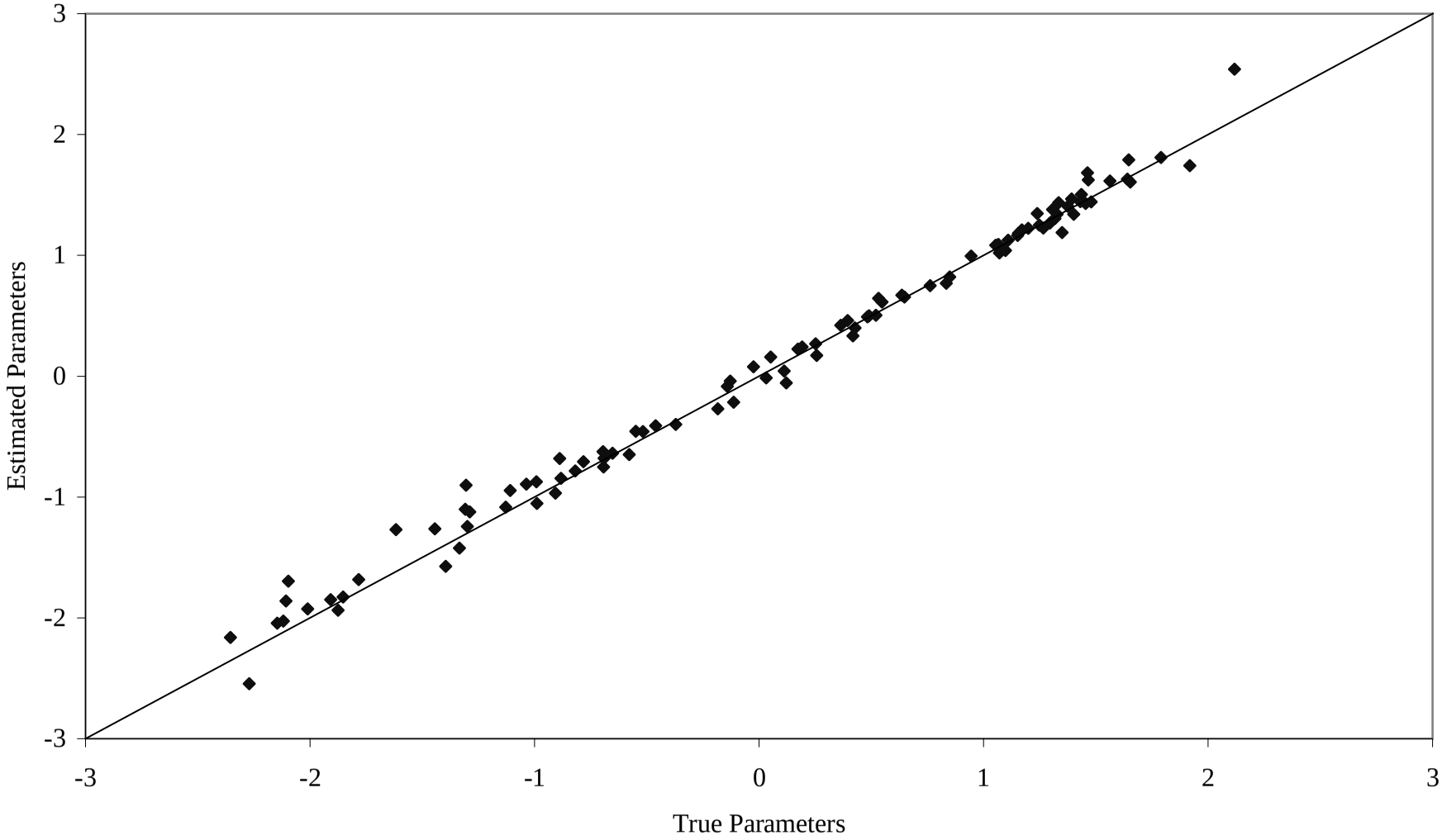
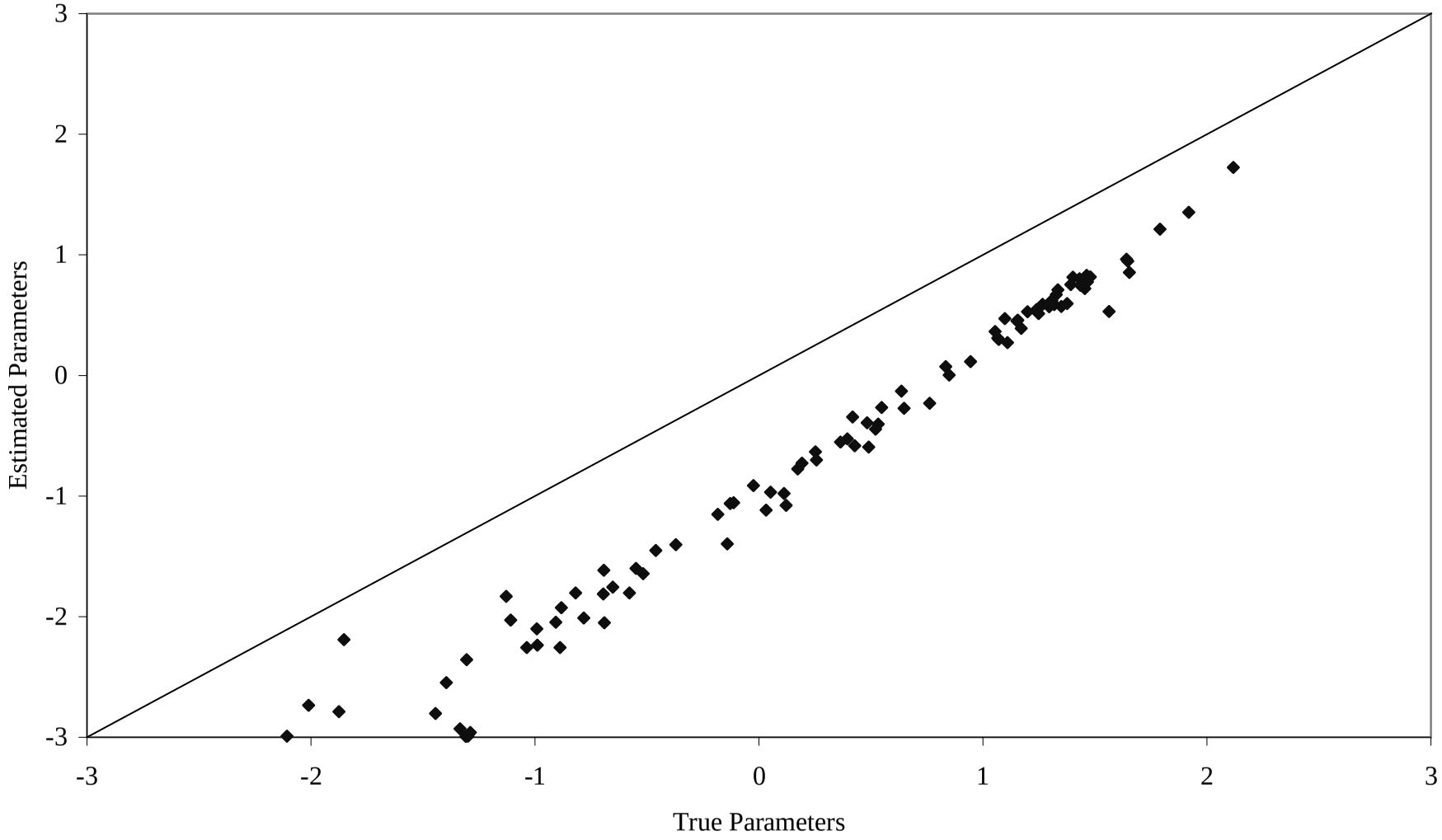


Figure 13. Recovery of b Parameters for 100 Pretest Items from a Bilog-MG Calibration Fixing 137 Parameters and Estimating 468 Parameters, for CAT Data Simulated from a N(+1,0.8) Distribution.




Note: Two items with estimated b parameters of < -7.0 are not pictured here.

Figure 14. Recovery of b Parameters for 100 Pretest Items from a Bilog-MG Calibration Fixing 137 Parameters and Estimating 468 Parameters, for CAT Data Simulated from a N(-1,1.2) Distribution.

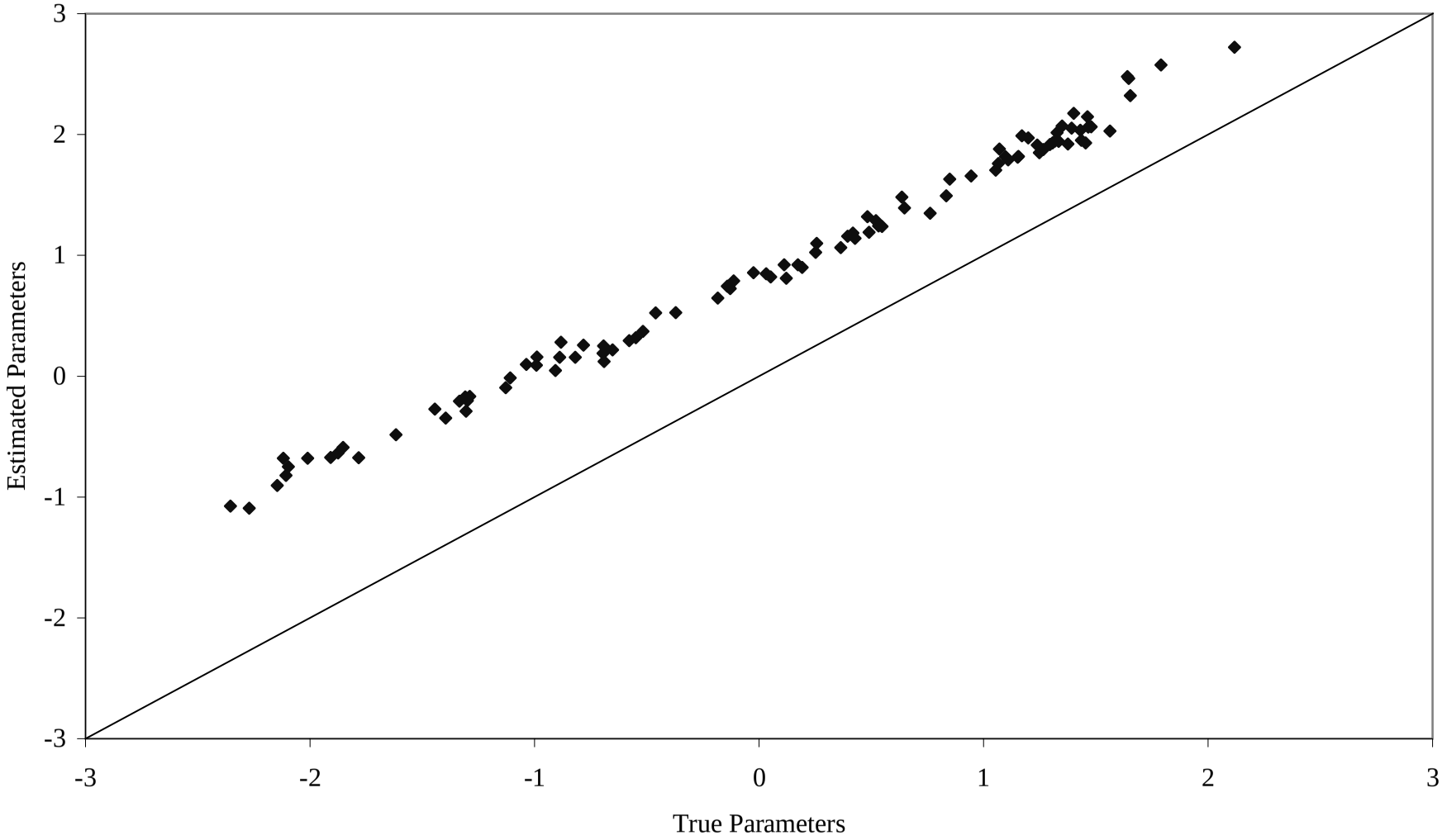
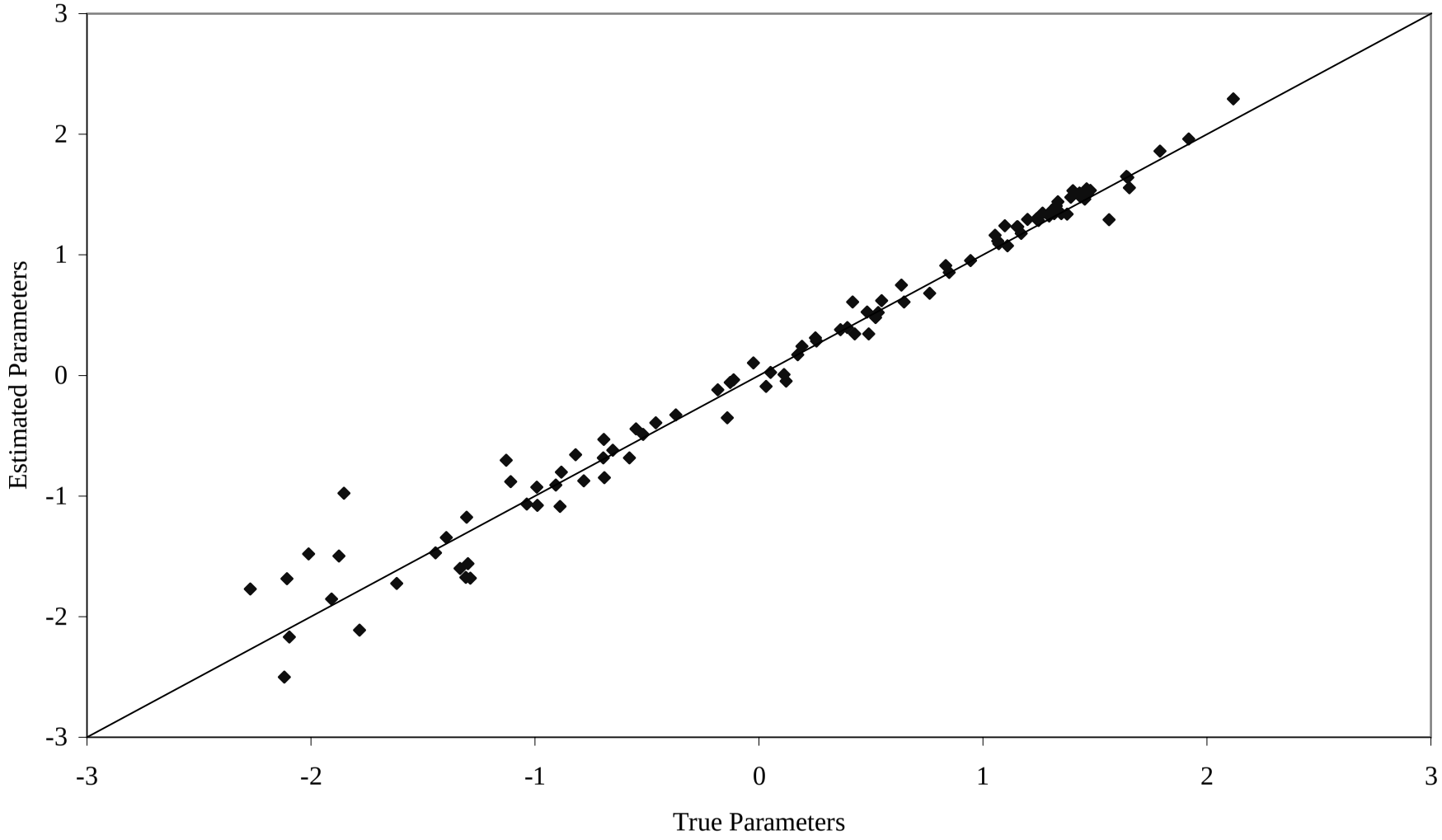


Figure 15. Recovery of b Parameters for 100 Pretest Items When the MML Transformation Procedure was Used to Rescale the Parameters, for CAT Data Simulated from a N(+1,0.8) Distribution.




Note: Two items with estimated b parameters < -5.0 are not pictured here.

Figure 16. Recovery of b Parameters for 100 Pretest Items When the MML Transformation Procedure was Used to Rescale the Parameters, for CAT Data Simulated from a N(-1,1.2) Distribution.

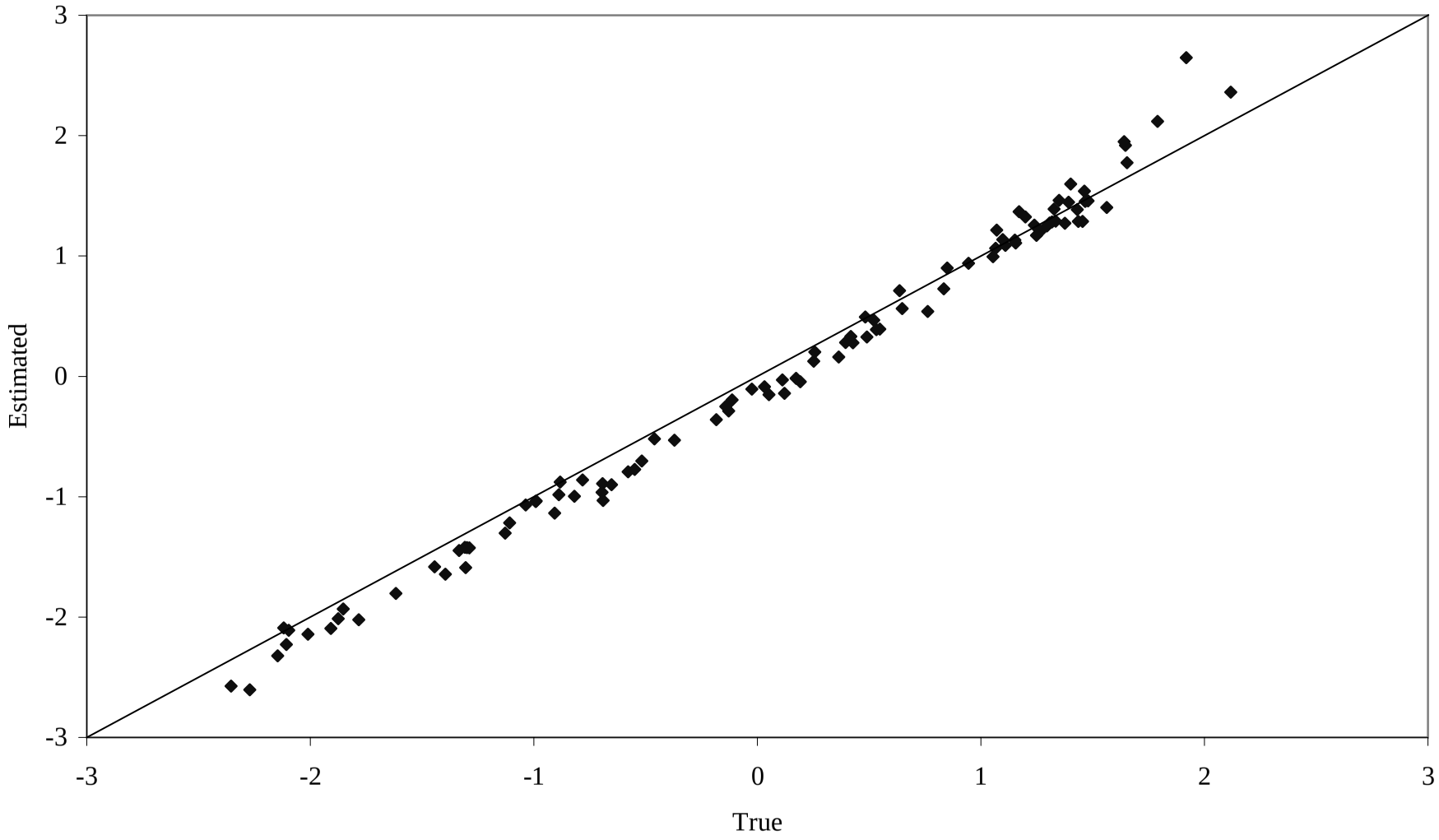


Figure 17. RMSD for 100 Pretest Items, for Parameters Estimated Using the Bilog-MG Fix Capability and Parameters that are Rescaled Using the MML Transformation Procedure, for CAT Data Simulated from a N(+1,0.8) Distribution.

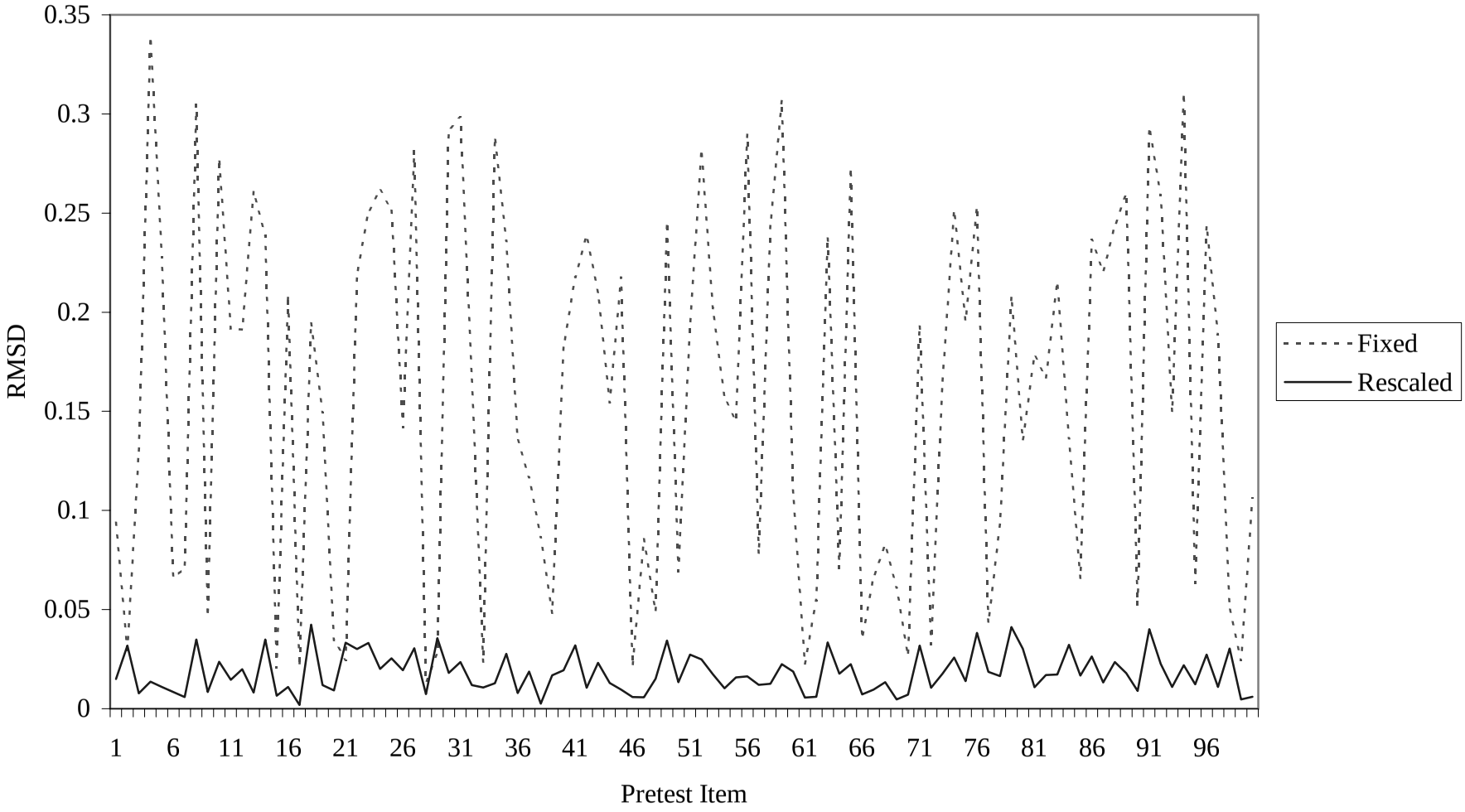


Figure 18. RMSD for 100 Pretest Items, for Parameters Estimated Using the Bilog-MG Fix Capability and Parameters that are Rescaled Using the MML Transformation Procedure, for CAT Data Simulated from a N(-1,1.2) Distribution.

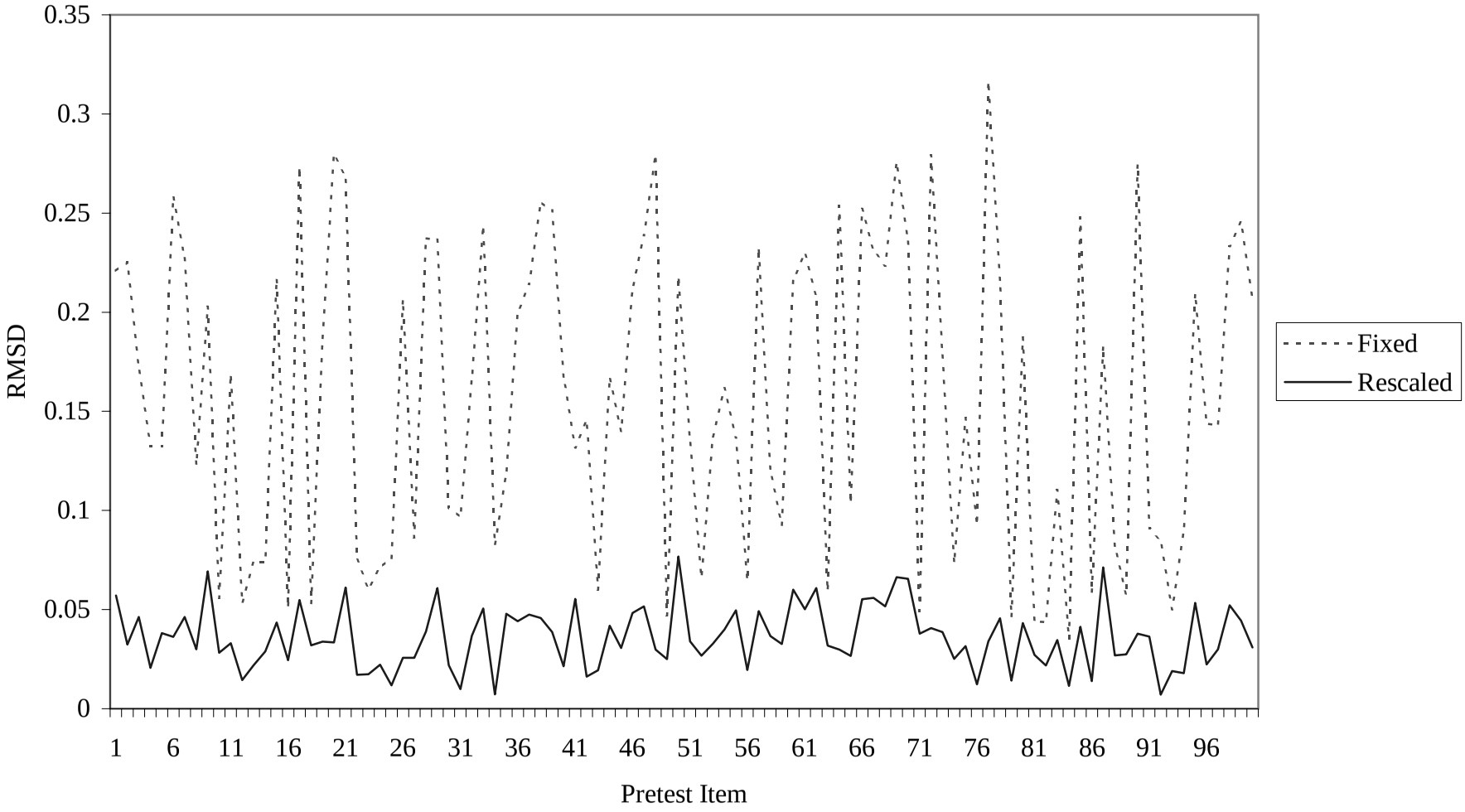


Figure 19. RMSD for 100 Pretest Items, for Parameters Estimated Using the Bilog-MG Fix Capability and Parameters that are Rescaled Using the MML Transformation Procedure, for CAT Data Simulated from a N(0,1) Distribution.