Maintaining Scale in Computer-Adaptive Testing

Robert L. Smith
Saba Rizavi

Roxanna Paez

Michele Damiano

Erin Herbert

Educational Testing Service, Princeton, New Jersey

Paper presented at the annual meeting of the American Educational Research Association (AERA) and the National Council on Measurement in Education (NCME) held between
April 21 to 25, 2003, in Chicago, IL.

Unpublished Work Copyright © 2003 by Educational Testing Service. All Rights Reserved.

These materials are an unpublished, proprietary work of ETS. Any limited distribution shall not constitute publication. This work may not be reproduced or distributed to third parties without ETS's prior written consent. Submit all requests through www.ets.org/legal/copyright.

Maintaining Scale in Computer-Adaptive Testing[1]

One of the most important considerations in a paper-and-pencil testing program is making sure that test scores are comparable over time, i.e., the scale does not drift. In a computer-adaptive environment (CAT) using item response theory (IRT), the analogue is that the items are placed on the same underlying scale so that the adaptive test scores produced are comparable, i.e. on the same scale.

In traditional applications of IRT methods with paper-and-pencil tests, the accepted procedure is to independently calibrate items on different forms, then to scale these items by linearly transforming the abilities to a base scale through the test characteristic curves for a set of common items (Stocking and Lord, 1983). This method was developed using relatively large samples to obtain stable parameter estimates. Many operational CAT programs have continued to scale pretest items using this two-step process with much smaller samples.

In a CBT environment where pretest slots are valuable, there has been interest in finding scaling methods that would not "waste" valuable pretest slots with linking sets only used for scaling. One such method is to use item-specific priors (ISP) on the operational items to hold the scale. Here the priors are based on the item parameter estimates and the item parameter variance-covariance matrix obtained from the pretest calibration. Items that are better estimated (i.e., with smaller standard errors) or that have been estimated using larger samples have stronger priors and contribute to maintaining the scale to a greater degree. With the ISP method the item parameters are treated as

[1] We would like to thank Charlie Lewis for many thoughtful discussions that have improved the paper appreciably.

estimates, not totally known, and are able to move to accommodate additional information from the data.

Theoretically, the scaling through an anchor should be unnecessary with the ISP method. However, on-line calibration studies (e.g., Folk & Golub-Smith, 1996) have been inconclusive as to whether the scaling using the common items provides benefit or introduces error into the scaling.

A possible reason for the inconclusive results using this method may be the "strength" of the item priors used to hold the scale. Paper-and-pencil calibrations were often based on a thousand or more test takers resulting in relatively well-estimated item parameters and strong item priors. Target samples for CAT programs using PARSCALE are often as low as 500 test takers. The smaller sample size is likely to result in item parameters that are less well estimated and with item priors that are easily "over-powered" by the likelihood based on operational CAT data from thousands of test takers. The ability of items calibrated on small samples (with diffuse posteriors) to hold the scale is questionable.

Ban, Hanson, Wang, Yi, and Harris (2001) recently examined five on-line calibration methods. These included: (1) fixing abilities while estimating pretest item parameters using joint maximum likelihood methods (Stocking Method A, Stocking, 1988); (2) fixing abilities while estimating pretest item parameters using joint maximum likelihood methods, then linearly scaling the abilities through the test characteristic curve for the linking items (Stocking Method B, Stocking and Lord, 1983; Stocking, 1988); (3) a marginal maximum likelihood method with a single EM cycle where in the E-step the posterior distribution for ability is based only on the operational items and in the M-step,

the operational item parameters are fixed while the pretest items are updated (MMLE/OEM, Wainer & Mislevy, 1990); (4) a marginal maximum likelihood method with multiple-EM cycles where in the E-step the posterior distribution is based on both operational and pretest items, and in the M-step the operational item parameters are fixed while the pretest items are updated (MMLE/MEM, Ban, Hanson, Wang, Yi, & Harris, 2001); (5) a marginal maximum likelihood estimation method with strong Bayesian priors on the operational item parameters (BILOG/PARSCALE, Mislevy & Bock, 1990; Muraki & Bock, 1995). With this method the item parameters are essentially fixed by making the standard errors extremely small.

The Stocking B, MMLE/MEM and BILOG methods performed comparably in the Ban, Hanson, Wang, Yi, & Harris (2001) study. However, the authors favored the MMLE/MEM method because it did not require a set of linking items, and thus, would not use up valuable pretest slots, as does Stocking's Method B. This method also worked with small to moderate sample sizes, where the BILOG method employed in the study only converged when very large sample sizes (N=3000) were used. One reason the BILOG method may not have worked for smaller sample sizes is that the priors used were too tight or too discrepant from the likelihood to produce a solution. For all of the methods examined by Ban, Hanson, Wang, Yi, and Harris (2001) the item parameters for the operational items were fixed (or essentially fixed). However the fixing of parameter estimates to maintain the scale may introduce a drift in the scale since the parameter estimates contain error, but are treated as true parameter values.

With the exception of the Stocking (1988) study, all studies examined item parameter recovery to assess the viability of the methods. Item parameter recovery only

indirectly addresses whether the scale has been maintained. A more direct assessment of scale drift is to examine cycles of calibration to see if there is a cumulative effect from the different scaling methods and to evaluate the drift in terms of whether scores are impacted by any observed drift. The present study examined scale drift in a CAT environment across multiple pretest cycles with operational item constraints and exposure control. Three direct scaling methods, (1) Item-Specific Prior (ISP), (2) Fixed parameter estimates (Fixed, a variation of MMLE/MEM), and (3) a compromise method where ISP standard errors are "shrunken" to produce stronger priors without fixing the operational parameter estimates (Shrunken). Scaling through an anchor was also examined for each direct scaling method.

Method

Data

Four Quantitative Reasoning operational pools from a high-stakes CAT program, constructed for operational use, were used in the study. All items were discrete to avoid set dependencies. All CATs contained 28 items. A 28-item anchor, external to the operational pools, was available to be used for scaling (similar to the Stocking B method). All items were calibrated using a 3-parameter logistic model. The true score scale is based on a reference test with 64 items.

Pretests

Item pools were used both for the "operational" CATs and as pretests. When the first pool was used for the operational CAT, the second pool provided the items from

which to construct the pretests. Each item pool contained 395 items. Fourteen "pretests" of 28 items[2] each were formed from each item pool. Pretests were assembled to an average information function for a pool and thus were approximately statistically parallel. Content information was not used in the assembly of the pretests. With the inclusion of the anchor, there were 15 pretests. The anchor was included in all calibrations. However, in cases where scaling did not use the anchor, the calibration estimates were ignored. Pretest sample sizes of 500 and 1000 test takers were used.

Calibration methods

Three calibration methods (Bayesian item specific prior, MMLE/MEM with fixed operational parameter estimates, shrunken ISP) and two anchor-scaling conditions (direct scaling and scaling through an anchor test characteristic curve) were crossed to produce six calibration/scaling (e.g., item specific prior method with anchor scaling) conditions.

Item Specific Prior (ISP) method

The item specific prior method (ISP) uses a multivariate normal form as the prior for each set of item parameters with item parameter prior means, standard errors, and the item parameter variance-covariance matrix for an item to define the prior. The item priors for an operational pool are used to define and hold the scale while the pretest items are calibrated and placed directly on scale. Data from both the operational and pretest items are used during estimation. As suggested above, items with stronger priors and

[2] Each pool contained 395 items. Fourteen 28-item pretests can be constructed from a pool , however, three items will be left unassigned to a pretest. In order to have complete pools in each stage, three pretests were constructed with 29 items.

smaller standard errors make larger contributions to holding the scale than more poorly estimated items with diffuse item priors.

Shrunken Item Specific Prior (SISP) method

The priors used in the ISP method are based on the sample size at the time of calibration. To maximize the number of pretests that may be calibrated, the estimation sample size may be as low as 500 test takers. A concern is that the priors based on such small samples may be insufficient to hold the scale in a CAT environment where some items are used much more often than others even when conditional exposure control is used. For operational items that receive a large number of exposures the priors may be "over-powered" by the data, which may introduce scale drift. However, if priors were strengthened, the scale might better be maintained.

In order to investigate whether strengthening the prior would have an effect on the maintenance of the scale, the item parameter variance-covariance terms for each item were shrunken by a factor of .25. This resulted in the standard errors being reduced by half. This method is a middle ground between fixing the item parameter estimates and using ISPs from the calibration.

Marginal maximum likelihood with multiple E-M cycles (MMLE/MEM)

The MMLE/MEM method used by Ban, Hanson, Wang, Yi, and Harris (2001) fixed the operational item parameters, then estimated the posterior theta distributions based on the responses to the operational items. They then used the estimated posterior theta distributions to estimate item parameters for the pretest items only, keeping the

operational item parameters fixed. The MEM method continues with E-M steps until pretest item parameter estimates converge. Each E-step (after the first) uses all item responses to update the posterior theta distributions.

The modified MMLE/MEM method used here also fixes the operational items at their calibrated estimates and continues until the pretest item parameter estimates converge. However, on the first as well as for all subsequent E-steps, all item responses are used to estimate the posterior theta distributions. Thus, in the first E-step, starting values for pretest item parameter estimates are used together with fixed operational item parameter estimates to estimate posterior theta distributions. This is expected to have little influence on the final item parameter estimates in relation to the MEM method used by Ban, Hanson, Wang, Yi, and Harris (2001).

Anchor Scaling

Each of the estimation methods was either used directly (direct scaling) or scaled using an anchor and the test characteristic curve (TCC) method (Stocking and Lord, 1983). Six scaling methods were examined in all: (1) IPS-Direct, (2) Shrunken-Direct, (3) Fixed-Direct, (4) ISP plus TCC scaling, (5) Shrunken plus TCC scaling, and (6) Fixed plus TCC scaling. It is important to recognize that each TCC scaling method constituted a separate, unique chain of scalings from their direct scaling counterparts.

Procedures

The study can be broken into two cyclical phases, a calibration phase and an evaluation phase. In the first calibration phase, scored item responses were generated for

3000 simulated test takers for the entire first pool and the anchor, as if, the entire pool, plus the anchor, had been taken by each simulated test taker. The first pool and the anchor were then concurrently calibrated using marginal maximum likelihood methods. The resulting estimates were placed on scale with the true item parameters using the anchor and the test characteristic curve (TCC) method (Stocking and Lord, 1983).

In the evaluation phase of the first cycle, only the first pool was administered as a CAT at 41 fixed true ability levels. The CAT data, in this phase, were generated using a three-parameter logistic model for 1200 simulated test takers at each of 41 true ability levels (49200 simulated test takers in all). Items were selected using the weighted deviations model (Stocking & Swanson, 1993) and a multinomial item exposure control procedure (Stocking & Lewis, 1998). Items were selected for delivery and abilities were estimated using the estimated item parameters obtained from the initial (base) item calibration. A probability table based on the true item parameters obtained from the operational vat was used to determine whether the item responses for a given simulated test taker was correct or incorrect.

The second calibration phase required the construction of a sparse matrix composed of CAT data, based on the first pool, and pretests composed of items from pool 2. Depending on the condition, either 500 or 1000 simulated test takers were independently sampled for each pretest. The pretest samples of test takers were generated based on probabilities for the test taking population at each of the 41 ability levels. This mimicked the sampling of test takers for a randomly administered pretest. For each pretest only one sample of 500 and 1000 test takers were generated. These samples were then used for all six scaling conditions.

Prior to item calibration in the second cycle, it was necessary to merge the pretest data with the operational data from the simulated CATs, matching on ability level. Counts were obtained at each ability level for the pretests by aggregating across all pretests and the anchor. Operational tests were randomly sampled from true ability strata and merged with the pretests based on the pretest counts at a given ability level. This resulted in a matrix composed of CAT items and pretest items matched on true ability and distributed according to the true distribution of ability. The resulting matrix was then used as input into the next calibration phase. All calibration methods were applied to this matrix at this stage.

In the third and subsequent cycles the processes were similar to stage two with the exception that all analyses were nested within a cell of a method x anchor design. For example, the pretests calibrated using ISPs (with no anchor scaling) in cycle 2 served as the operational pool item parameter estimates for the cycle 3 ISP (with no anchor scaling) simulations. These simulation results were then sampled down and joined with the pretests for pool 3. The resulting matrix was then calibrated using the ISP (no anchor scaling) method. This was done separately for each method x scaling x pretest sample size combination. In the final cycle, the pool 4 items were used to simulate the CAT data, while the pool 1 items served as the pretest items, thus completing a full chain.

Reference test. A reference test provided a transformation between θ and the number-right true score metric through a test characteristic curve (TCC). The reference test was an operational paper-and-pencil form of the test composed of 64 items. The item

parameters for this test were on the same scale as the other item parameters used in the study.

Evaluation criteria

For all methods, the full CAT simulation results at the final stage were used to assess the relationship between the number-right true scores and the estimated number-right true scores. To gain an overall assessment of the variation between the number-right true scores and the estimated number-right true scores for the different methods, root weighted mean-squared differences were computed between the number-right true score and the final estimated number-right true score across the ability scale. This is given by

$$RWMSD = \sqrt{\sum w_k Avg\left[\left(\tau_k - \hat{\tau}_k\right)^2\right]}$$

where w_k is the weight at a given true score,

τ_k is the number-right true score at one of the 41 abilities, and

$\hat{\tau}_k$ is an estimated number-right true score based on CAT responses for a single test taken at one of the 41 abilities.

The weight used is the proportion of the population at each of the 41 ability levels (Figure 1 shows the distribution of ability). As a consequence, this measure provides information about difference in the middle of the score scale where most candidates score. However, it down-weights discrepancies toward the ends of the scale. Since we are concerned with the whole of the scale, this measure may serve as a convenient one-number summary. However, it may not be sensitive to differences between true scores and estimated true scores near the ends of the scale.

Insert Figure 1 here

One way of capturing variance information across the scale is to compute conditional root mean-squared differences at each of the 41 points on the scale. Conditional values were computed using

$$RMSD_k = \sqrt{Avg\left[\left(\tau_k - \hat{\tau}_k\right)^2\right]},$$

The two measures just described provide information about total variation from true scores. The total mean squared difference includes two components: a random variance component like equating error and a squared bias component. Since we are concerned with scale drift, our primary focus is on the latter.

To obtain an assessment of bias, conditional mean differences were computed between the number-right true score and the final estimated number-right true score. Assuming random variation is evenly distributed around the true scores, positive and negative random error would tend to cancel each other, leaving the bias component. Directional differences from the true scores at different points along the scale would indicate the method-estimation bias accumulated over multiple chained estimation cycles. The bias was computed using

$$Bias_k = Avg\left[\left(\tau_k - \hat{\tau}_k\right)\right],$$

Results

One method of gaining an overall assessment of the different methods is by examining the root weighted mean square difference (RWMSD) between the true score and the estimated true score under each of the methods for the final pool in the chain. Table 1 shows this information for all methods for pretest samples of 500 and 1000. The three

methods show similar root weighted mean-squared differences (RWMSD) between true scores and estimated true scores with slightly larger values for N=500 sample and for the test characteristic curve (TCC) methods, as expected. From this data it appears that the methods that scaled through an anchor show larger amounts of scaling error.

Table 1
Root weighted mean-squared differences true scores and estimated true scores for all scaling methods

	Method					
	ISP		Shrunken		Fixed	
N	Direct Scaling	Scaling w/ TCC	Direct Scaling	Scaling w/ TCC	Direct Scaling	Scaling w/ TCC
1000	3.73	3.78	3.73	3.81	3.73	3.83
500	3.84	3.98	3.80	3.98	3.95	4.00

However, the RWMSD may not be the best indicator of differences between the methods investigated since it averages over all score points and places a majority of the weight in the middle of the score scale. Conditional analyses may provide more information about any differences that may exist among the methods.

Conditional analyses based on the scaling of the last pool in the chain

Two types of plots were examined to evaluate the different scaling methods. First plots of the conditional root mean squared differences (RMSD) between true scores and estimated true scores were examined to show deviation from the true scores regardless of the direction of the difference. This is a measure of the conditional variability of the different methods. In addition, conditional mean differences were examined to assess the direction of the differences (bias) for a given method.

Insert Figure 2 here

Conditional analyses by sample size

Variation from true scores (N=1000). Figure 2 shows the conditional RMSD based on number-right true score[3] for all methods when the pretest sample size was 1000. The curves are very similar at the middle and upper ends of the scale. At the upper end of the score scale the method based on fixed item parameters and directly scaled appears to have slightly larger RMSD values. At the low end of the scale, the methods using a test characteristic curve transformation appear to have lower RMSD values and also appear to approximate the base scaling based on a sample of 3000. The direct scaling methods appear to have larger RMSDs at the low end of the scale.

Conditional Bias (N=1000). Figure 3 shows the conditional mean bias (true score - estimated true score) for all methods when pretest sample size was 1000. A difference of 0 means that for a given true score, the average of the difference between the true score and the estimated true scores for the 1200 simulated test takers was zero or that the average estimated true score equaled the true score. Again, at the high end of the scale, the direct scaling method with fixed parameter estimates shows the largest difference. At the upper end of the scale all of the direct-scaling methods show larger deviation from true scores than do the TCC scaling methods. At the low end of the score scale the bias shifts direction, but again the TCC methods produce scores closer to the true scores than do the direct scaling methods.

[3] Note that although the labels are for number-right true scores, the actual metric is the θ metric. This allows us to see more differences in the ends of the score scale.

Insert Figures 3 and 4 here

Variation from true scores (N=500). The results when pretests were based on sample sizes of 500 are very similar to those found with calibration sample sizes of 1000. Figure 4 shows the RMSD conditional on true score for methods when pretest sample sizes were 500. At the middle and upper ends of the scale the curves are very similar, though they seem to be less similar than those based on pretest sample sizes of 1000. At the low end of the scale, the methods using a test characteristic curve transformation show lower RMSD values than do the direct scaling methods. Also the TCC methods are closer to the base scaling based on a sample of 3000 than are the direct scaling methods.

Insert Figure 5 here

Conditional Bias (N=500). Figure 5 shows the conditional mean bias (true score - estimated true score) for all methods when pretest sample sizes were 500 simulated test takers. At the low end of the scale the direct scaling methods show larger amounts of bias than the TCC methods. At the high end of the scale the direct scaling method with fixed item parameter estimates tends to show the most bias. Unexpectedly, the TCC scaling methods also show a high level of bias around a true score of 50. The TCC results for the N=500 sample for this last pool seem to be anomalous since this pattern is highly discrepant from the N=500 direct scalings and seems to be quite different from the pattern observed for the same methods for the pretest sample sizes of 1000. It may be related to peculiarities in the anchor test data since this is a common component for each of the TCC methods that

would be different from the 1000 pretest data and would not have been used for the direct scaling.

Insert Figures 6-9 here

Comparison of scaling methods

To get a clearer appraisal of the scaling methods themselves we examined their relative performance within sample size (500 vs. 1000) and scaling type (direct scaling vs. TCC) categories. Figures 6-9 show the RMSD for each of these combinations. The TCC methods based on the 1000 pretest sample appears to be the most consistent across scaling methods, with all of the methods closely approximating the base scaling results. The fixed-parameter direct scaling method appears to be more variable than the other methods and the ISP method appears to be slightly less-variable than the other methods in the direct 1000 and 500 pretest conditions.

Insert Figures 10-13 here

Conditional Bias. The conditional bias plots (Figures 10-13) show more pronounced differences among methods. For the direct scaling methods with a sample size of 1000 (Figure 10), the fixed method tended to produce larger amounts of bias than the other methods. Across most of the true-score scale the Item Specific Prior (ISP) method shows slightly less or comparable levels of bias when compared with the other two methods. This appears to be true regardless of scaling type (direct or using a TCC) or pretest sample size.

When a TCC scaling through an anchor is added to the direct methods (N=1000), the level of bias is less than if the anchor scaling is not used (Figure 11). This seems to suggest that additional improvement in bias can be obtained for the direct scaling methods with the addition of a TCC scaling. It should also be noted that the level of bias is less with a pretest sample size of 1000 vs. 500.

The Fixed scaling method tended to produce larger bias at the higher end of the scale when direct scaling was used and for the TCC method for a pretest sample of 500. There was little difference between methods for the TCC scaling with pretest samples of 1000. The Shrunken method produced larger bias at the low end of the scale (Figures 12 and 13). The ISP method showed less bias for most of the lower portion of the scale, except for the very lowest true scores. The TCC scalings for a pretest sample size of 500 showed very similar results for all methods, although all tended to show substantial amounts of bias. Finally, it should be noted that the degree of bias is substantially less for all TCC methods when compared to the comparable direct scaling method.

Discussion/Conclusions

A number of studies have examined on-line calibration methods and whether anchor scaling helps or hinders the maintenance of a scale. The results have been inconclusive. The present study sought to examine a number of scaling methods in a context where scaling error could accumulate over many item pools. The results suggest that direct scaling methods produce scales that show greater variation from true scores than do the TCC methods at the low end of the scale for pretest sample sizes of both 500 and 1000. The TCC methods also tend to better approximate the level of variation found in the base scaling than

do the direct scaling methods. With a pretest sample size of 500, the fixed-direct scaling method showed greater variation from true scores than the other direct scaling methods and the base scaling at the high end of the scale. For a pretest sample size of 500, all TCC methods also showed greater variation from true score than did the base scaling at the high end of the scale. However, with a pretest sample size of 1000, the TCC methods all tracked the base scaling well in terms of variation from true scores.

The TCC scaling methods also appear to produce scales with less bias on the number-right true score metric than direct scaling methods. When the TCC scalings were used in conjunction with any of the methods used for direct scaling, more stable results were obtained regardless of which direct scaling method was used. It also appears that for the scaling methods examined, more bias is accumulated when pretest sample sizes of 500 are used vs. pretest sample sizes of 1000.

The Item Specific Prior (ISP) method showed slightly less or comparable levels of bias when compared with the other two methods across most of the true score scale. This appeared to be true regardless of scaling type (direct or using a TCC) or pretest sample size. Fixing item parameters during estimation (using CAT data) tended to produce larger bias at the higher end of the scale when direct scaling was used and for the TCC method for a pretest sample of 500. The Shrunken method produced larger bias at the low end of the score scale with pretest sample sizes of 500. Few differences were observed between methods for the TCC scaling when pretest sample sizes of 1000 were used.

References

Ban, J-C., Hanson, B. A. , Wang, T., Yi, Q., & Harris, D. J. (2001). A comparative study of on-line pretest item-calibration/scaling methods in computerized adaptive testing. *Journal of Educational Measurement*, 38, 191-212.

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

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

Muraki, E. & Bock, R. D. (1995*). PARSCALE: Parameter scaling of rating data* (Version 2.2). Mooresville, IN: Scientific Software, Inc.

Stocking, M. L. (1988). Scale drift in on-line calibration (Research Report, RR-88-28-ONR). Princeton, N.J.: Educational Testing Service.

Stocking, M. L. & Lewis, C. (1998). Controlling item exposure conditional on ability in computerized adaptive testing. *Journal of Educational and Behavioral Statistics*, *23*, 57-75.

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

Stocking, M. L. & Swanson, L. (1993). A method for severely constrained item selection in adaptive testing. *Applied Psychological Measurement*, *17*, 277-292.

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

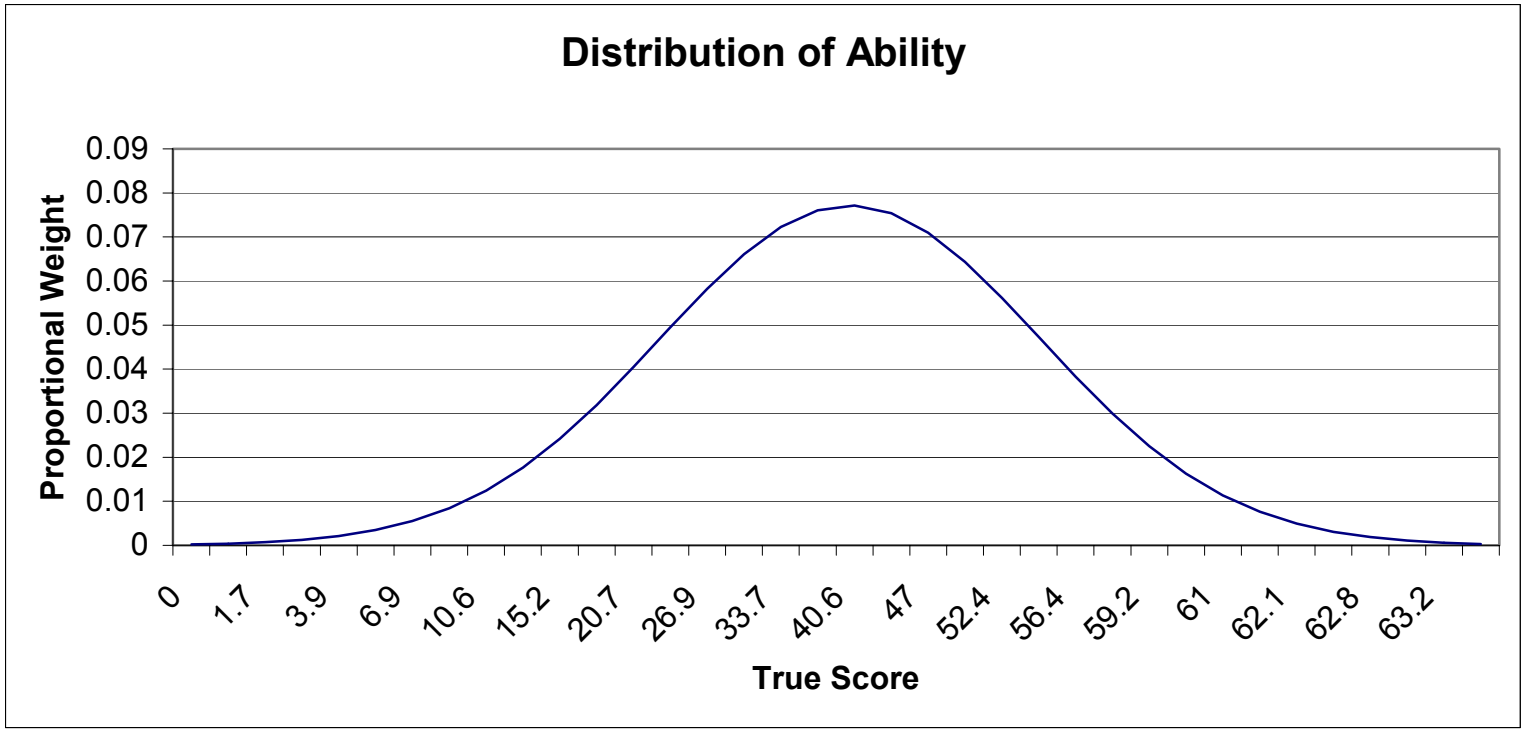

Figure 1
Distribution of ability

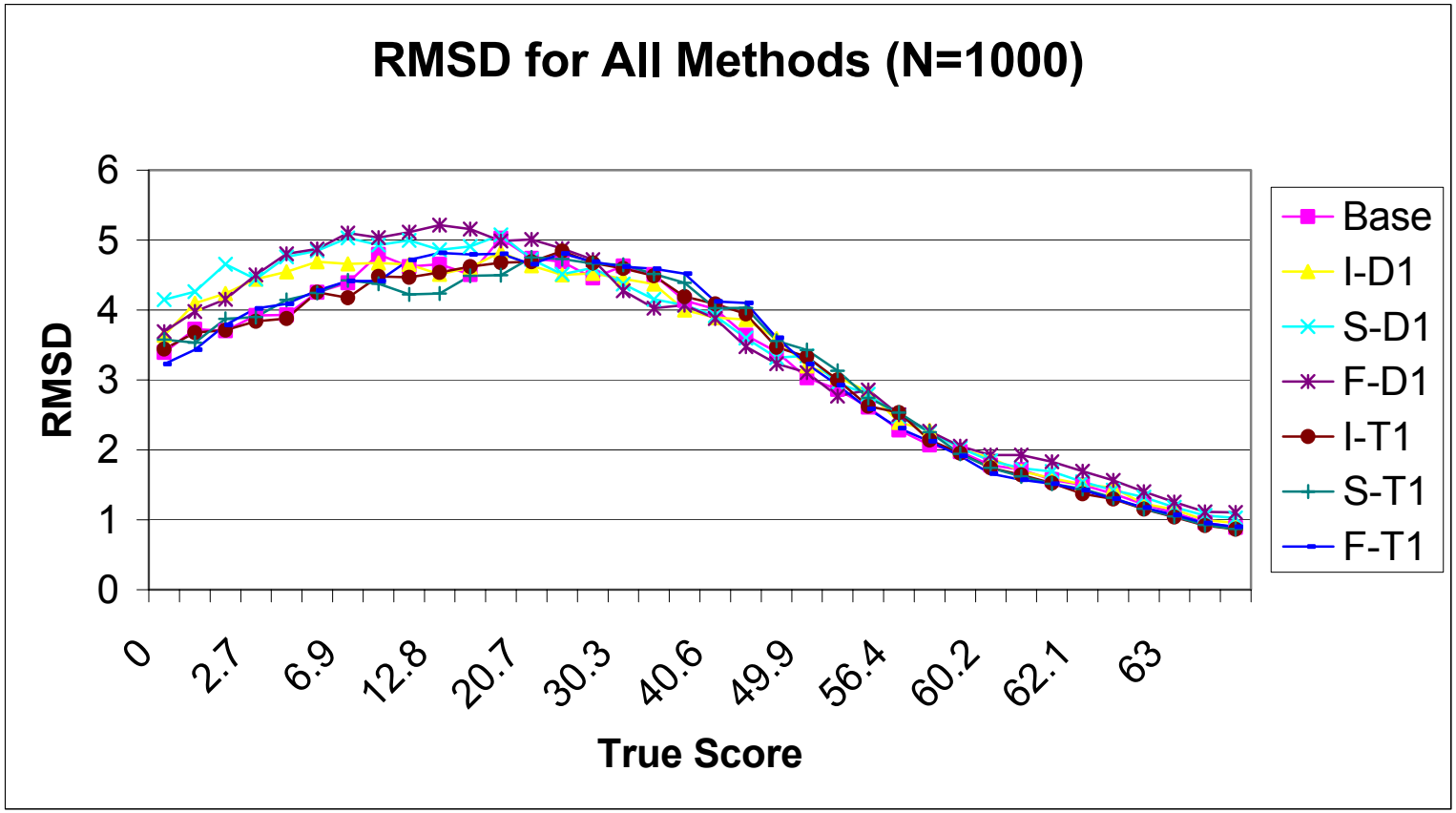

Figure 2[4]
Conditional variation from true scores for pretest sample sizes of 1000 for direct and TCC scaling methods

[4] The entries in the legends of the plots are coded as follows, Base=the initial base scaling, I-D1=ISP direct scaling based on a pretest sample of 1000, I-T1= ISP with TCC scaling based on a pretest sample of 1000. This is similar for the other scaling methods, where S=shrunken and F= fixed. A 5 in place of the 1 means that the pretest sample was based on 500 simulated test takers.

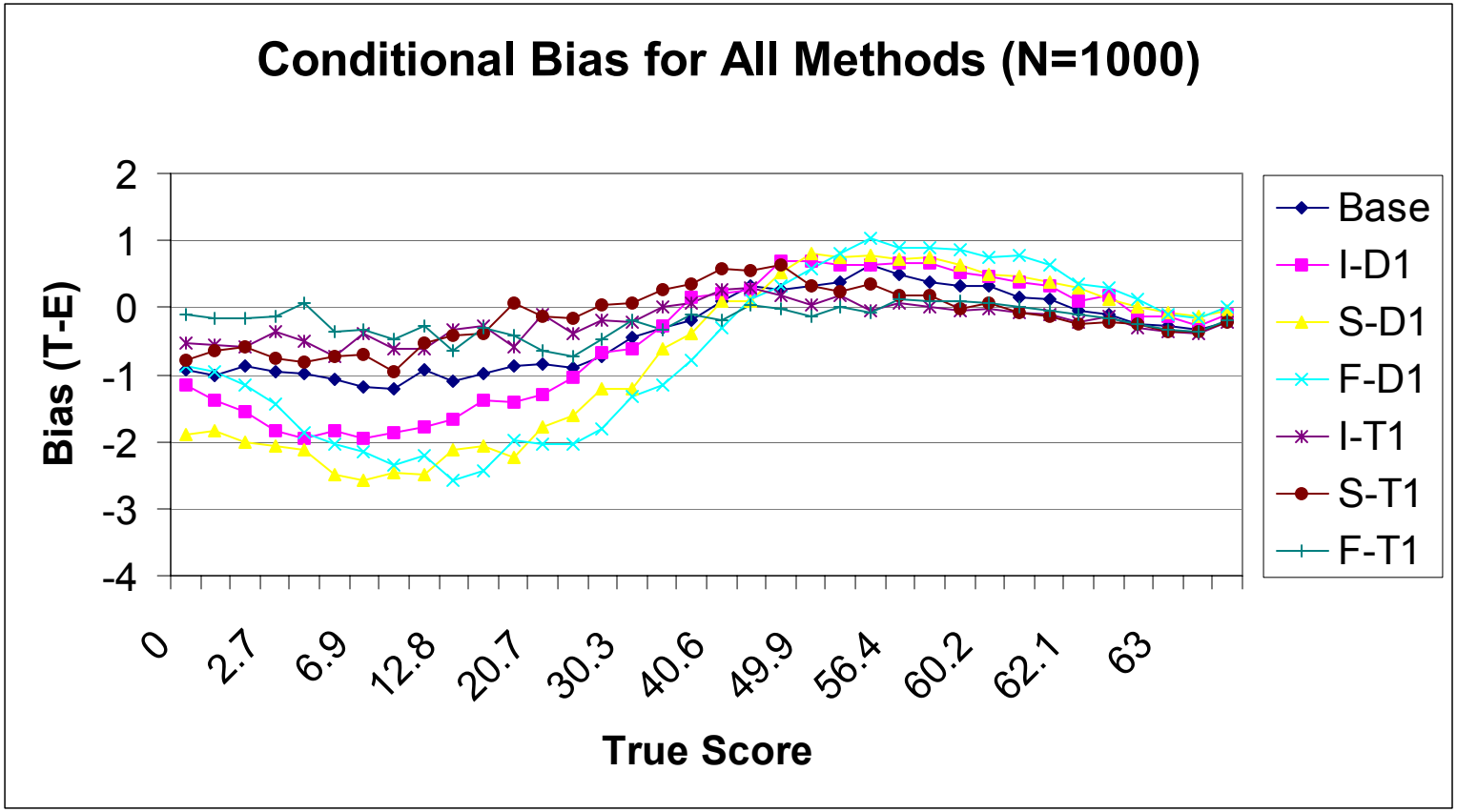

Figure 3
Conditional bias for direct and TCC scaling methods (N=1000)

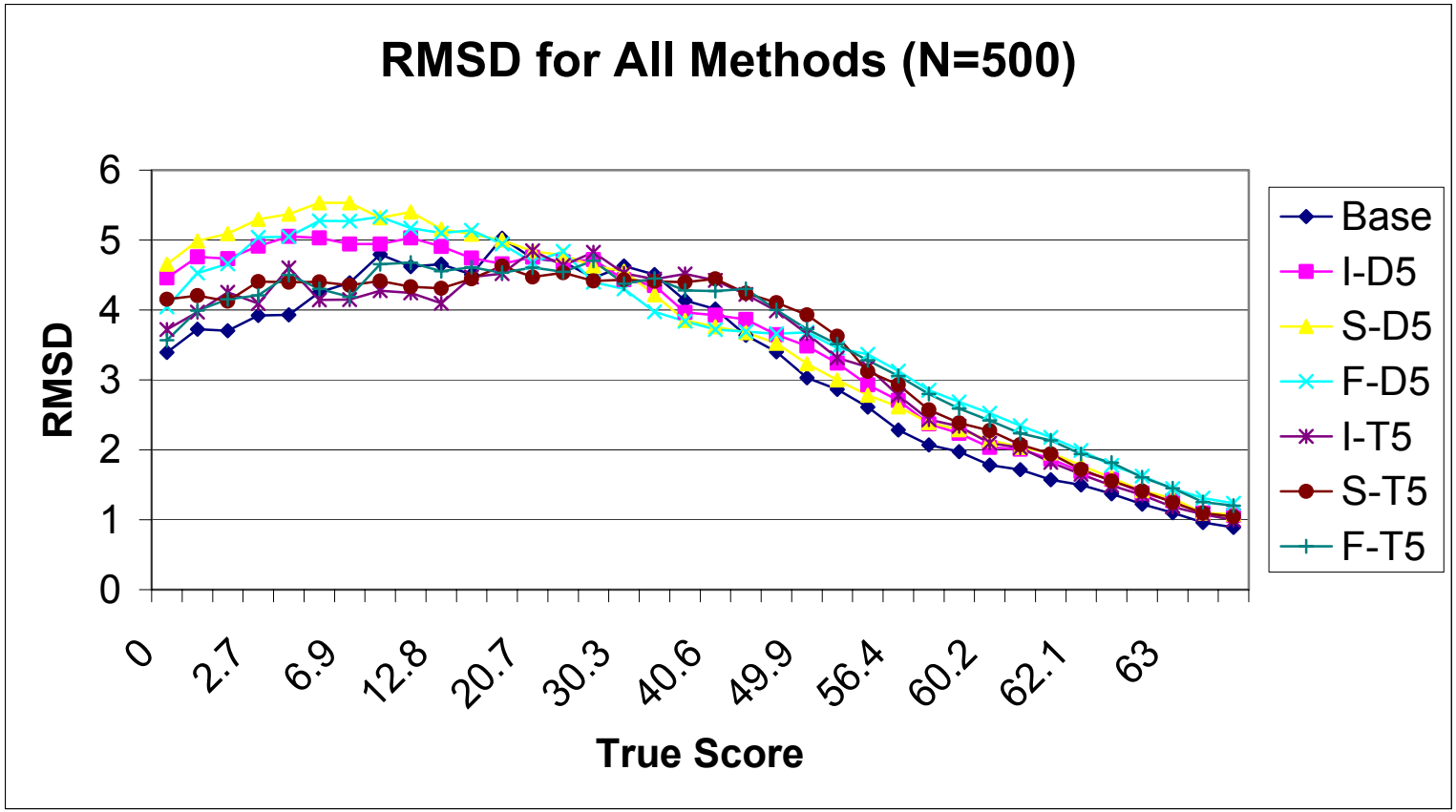

Figure 4
Conditional variation from true scores for pretest sample sizes of 500 for direct and TCC scaling methods

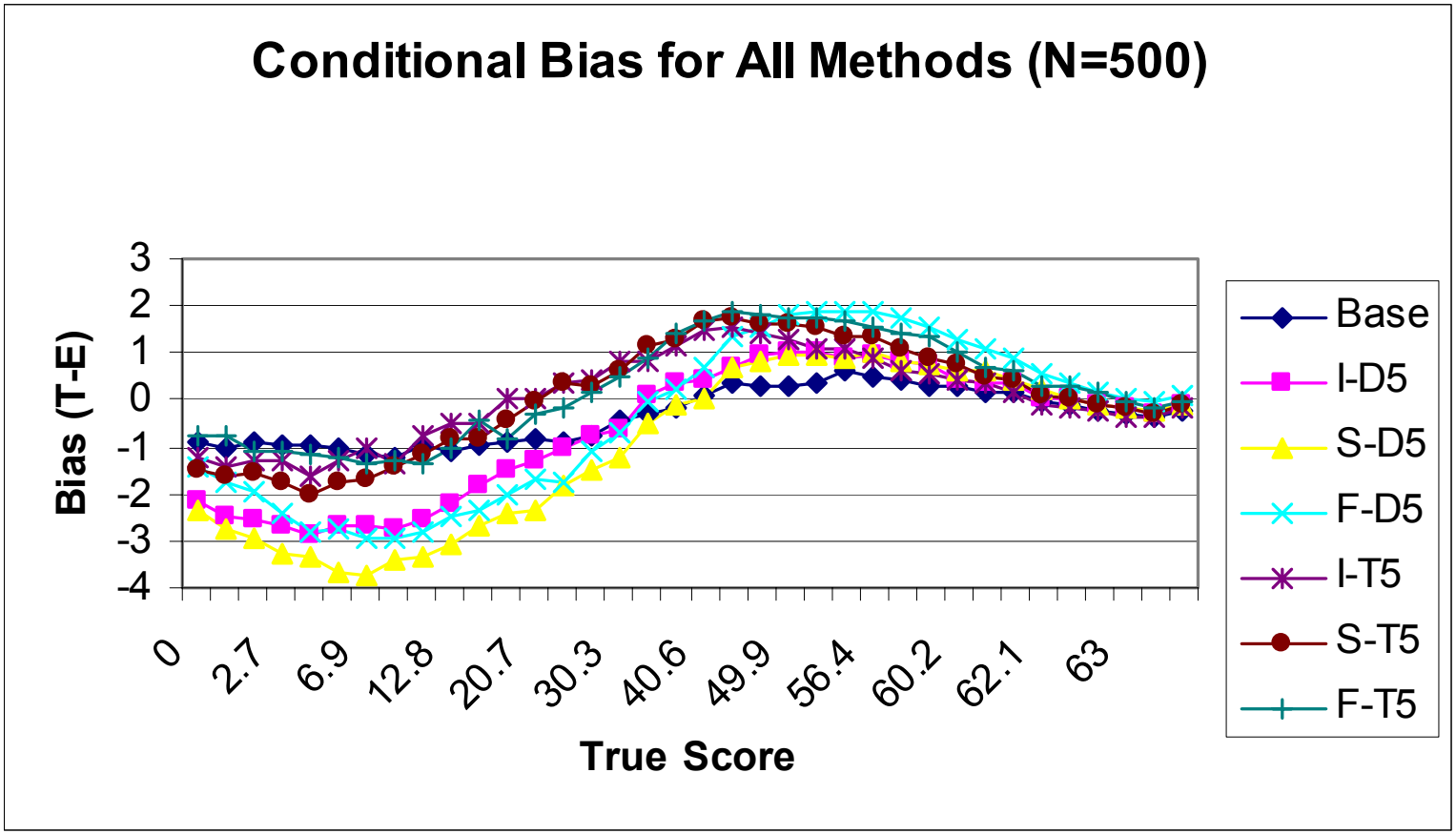

Figure 5
Conditional bias for direct and TCC scaling methods (N=500)

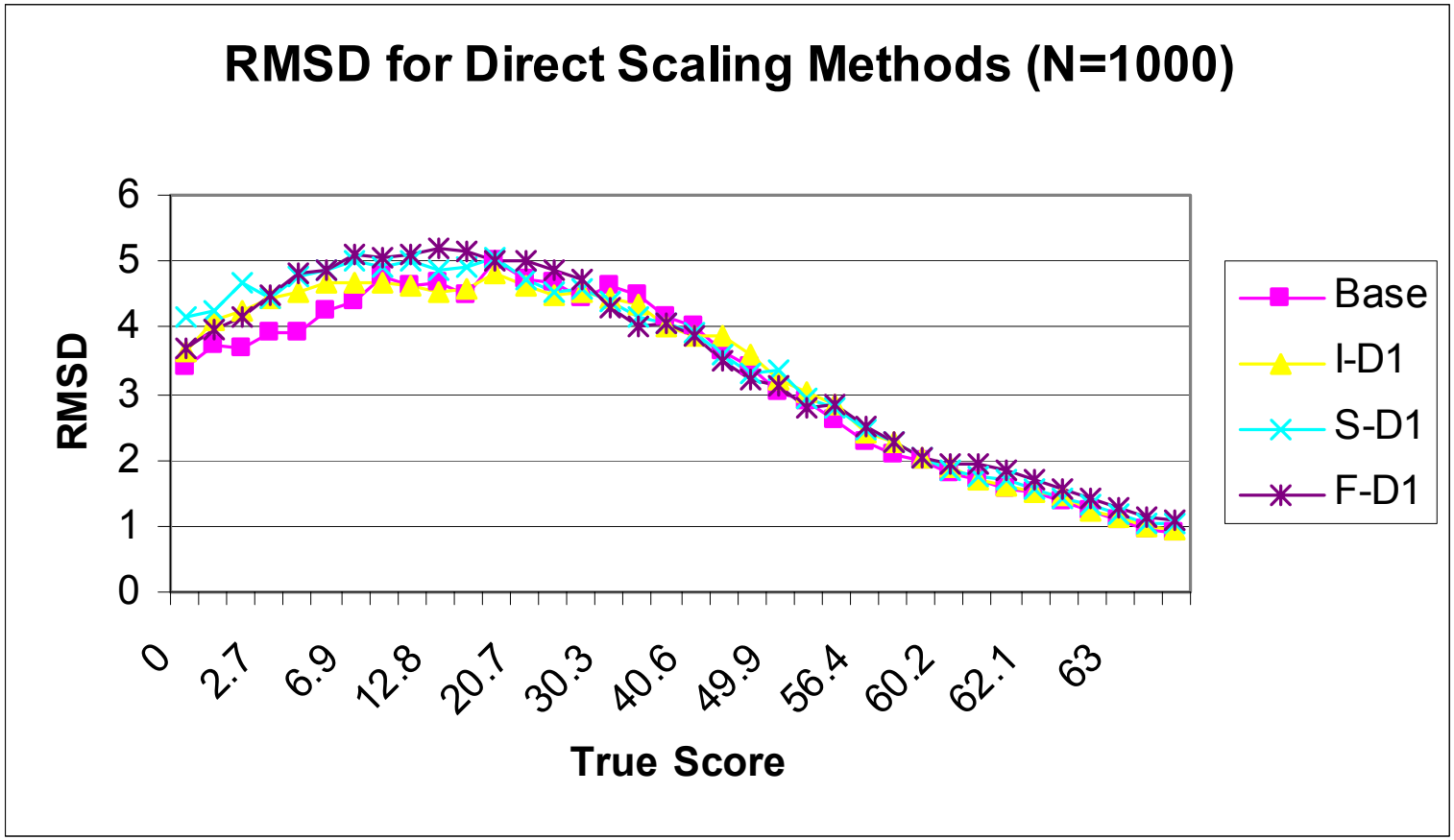

Figure 6
Conditional variation from true scores for pretest sample sizes of 1000 for direct scaling methods.

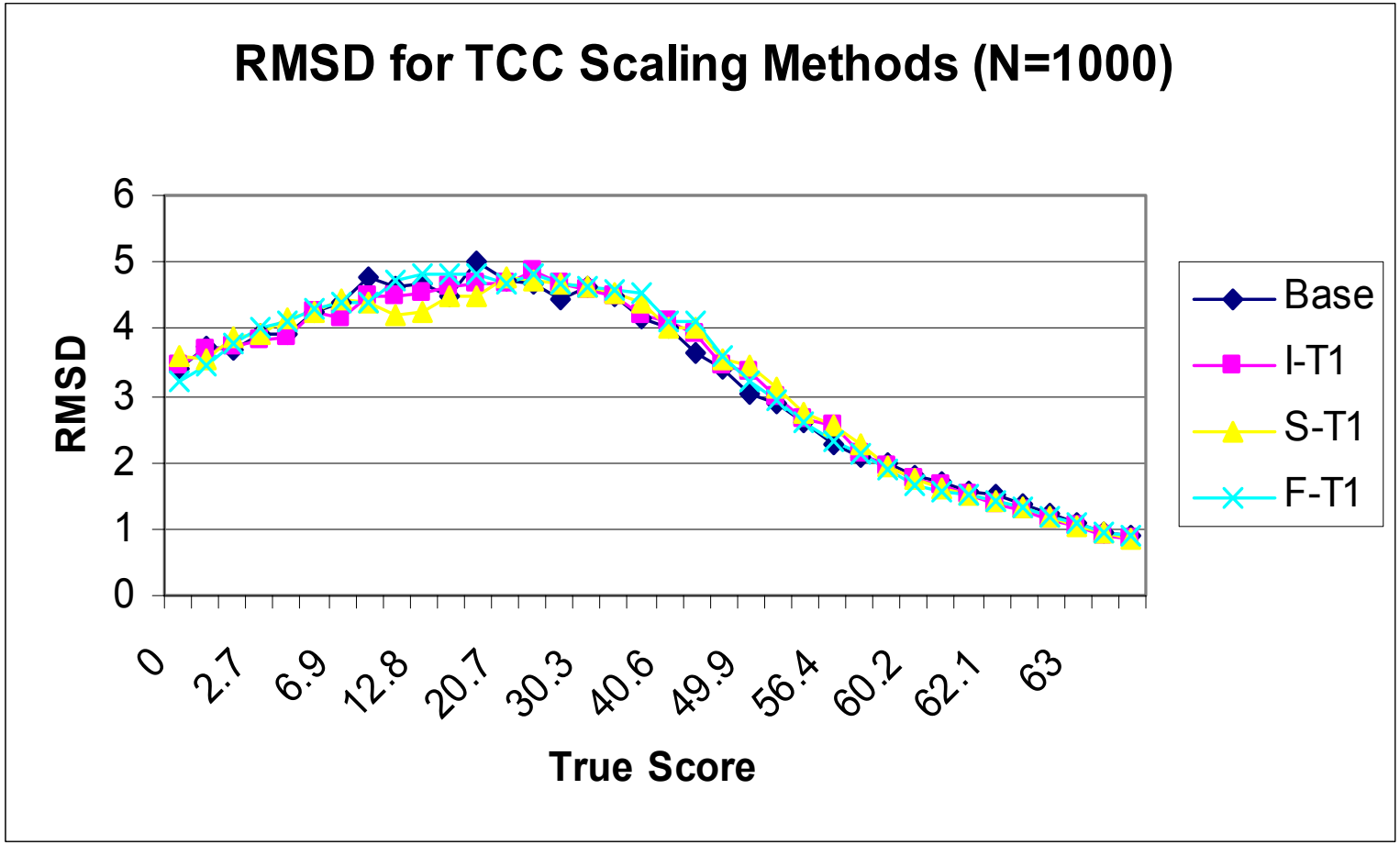

Figure 7
Conditional variation from true scores for pretest sample sizes of 1000 for TCC scaling methods

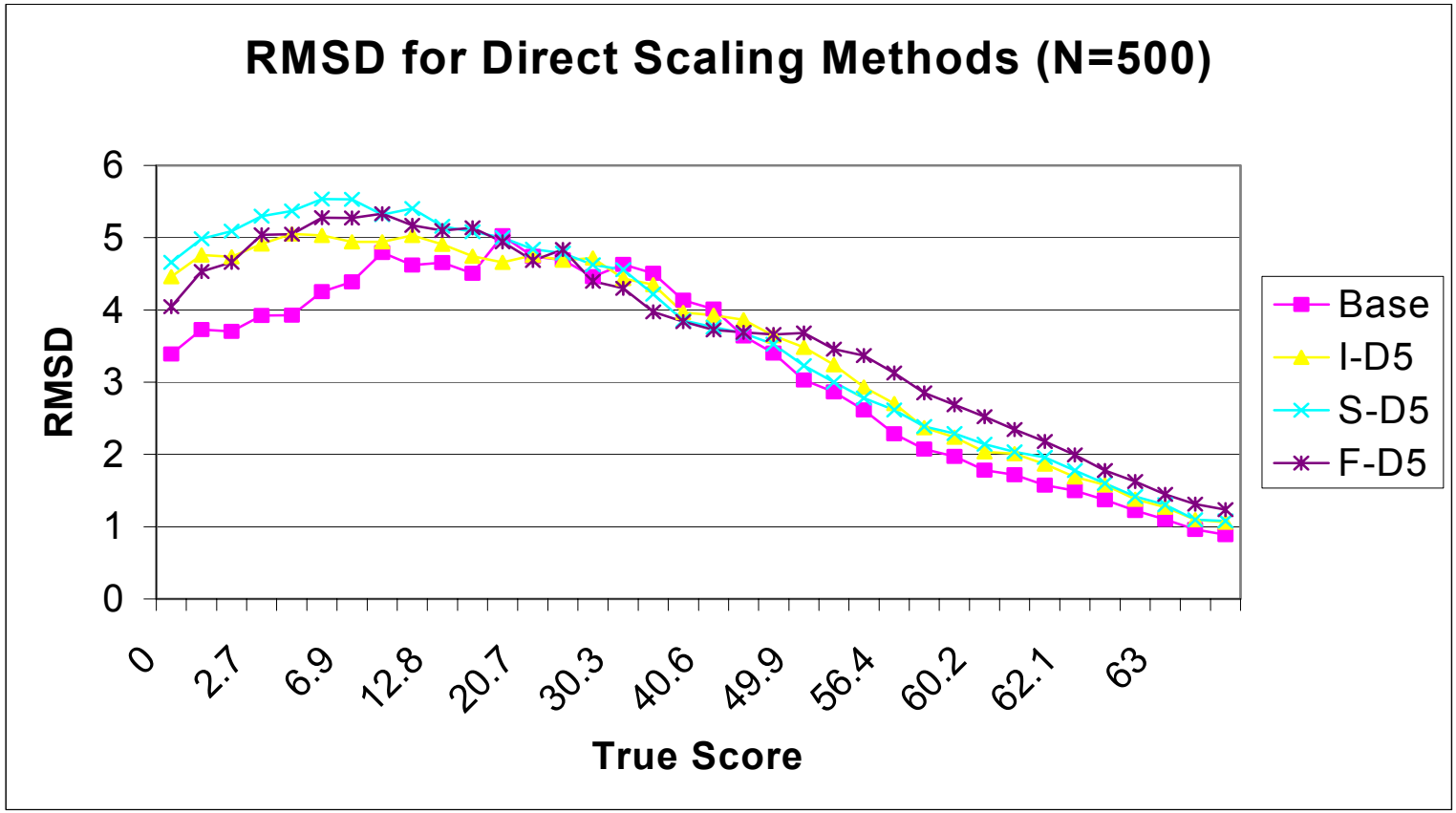

Figure 8
Conditional variation from true scores for pretest sample sizes of 500 for direct scaling methods

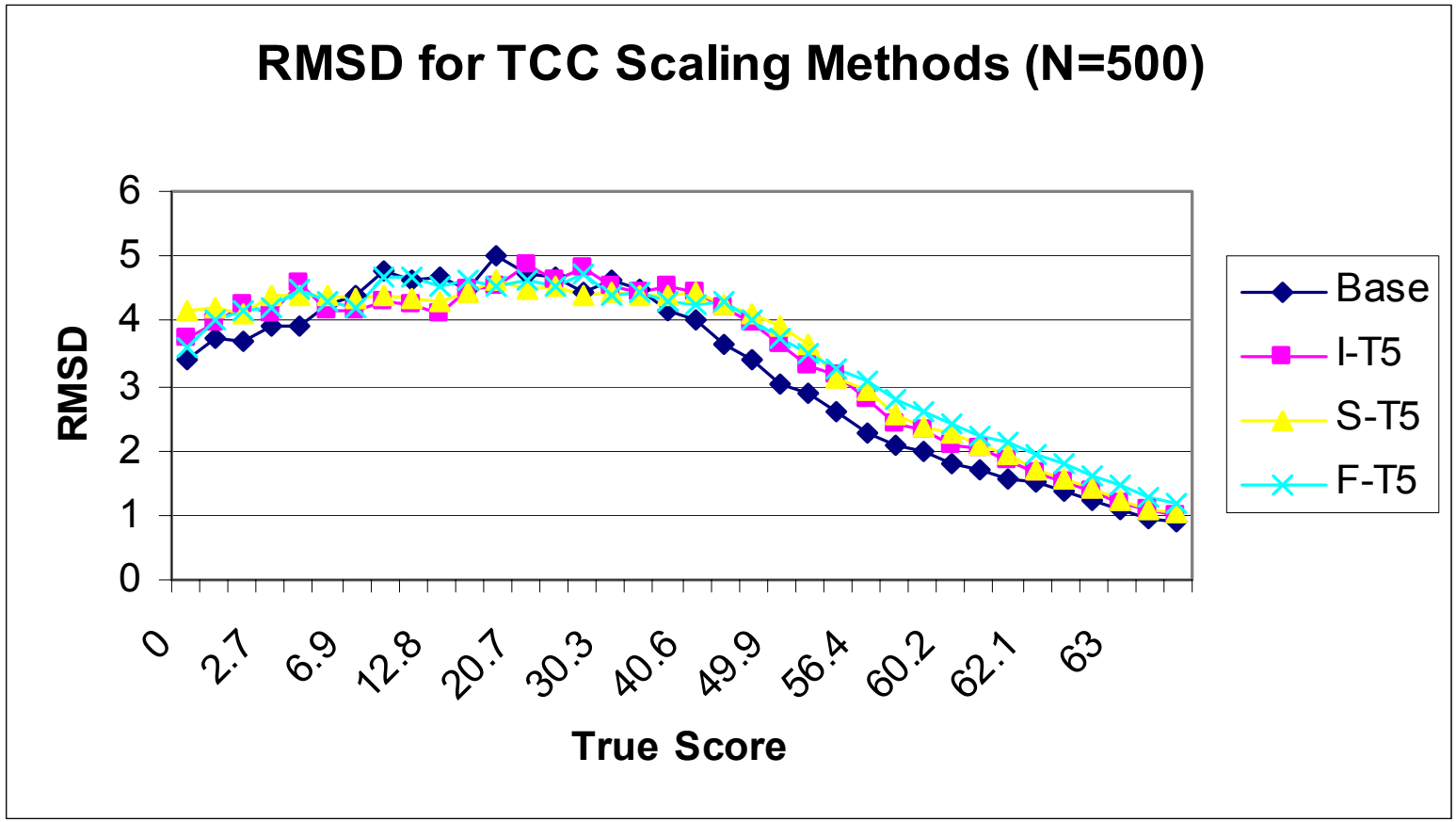

Figure 9
Conditional variation from true scores for pretest sample sizes of 500 for TCC scaling methods

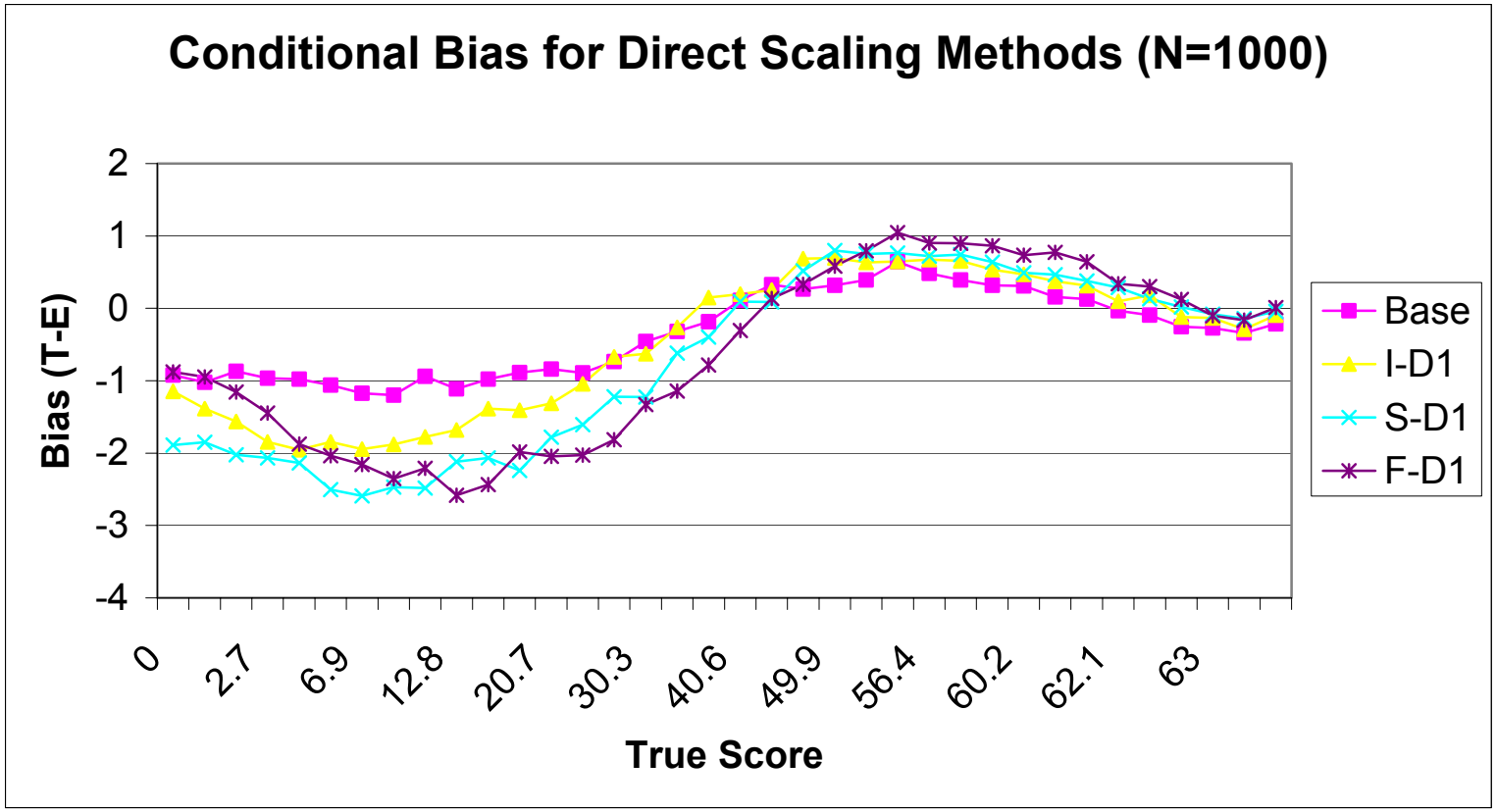

Figure 10
Conditional bias for direct scaling methods (N=1000)

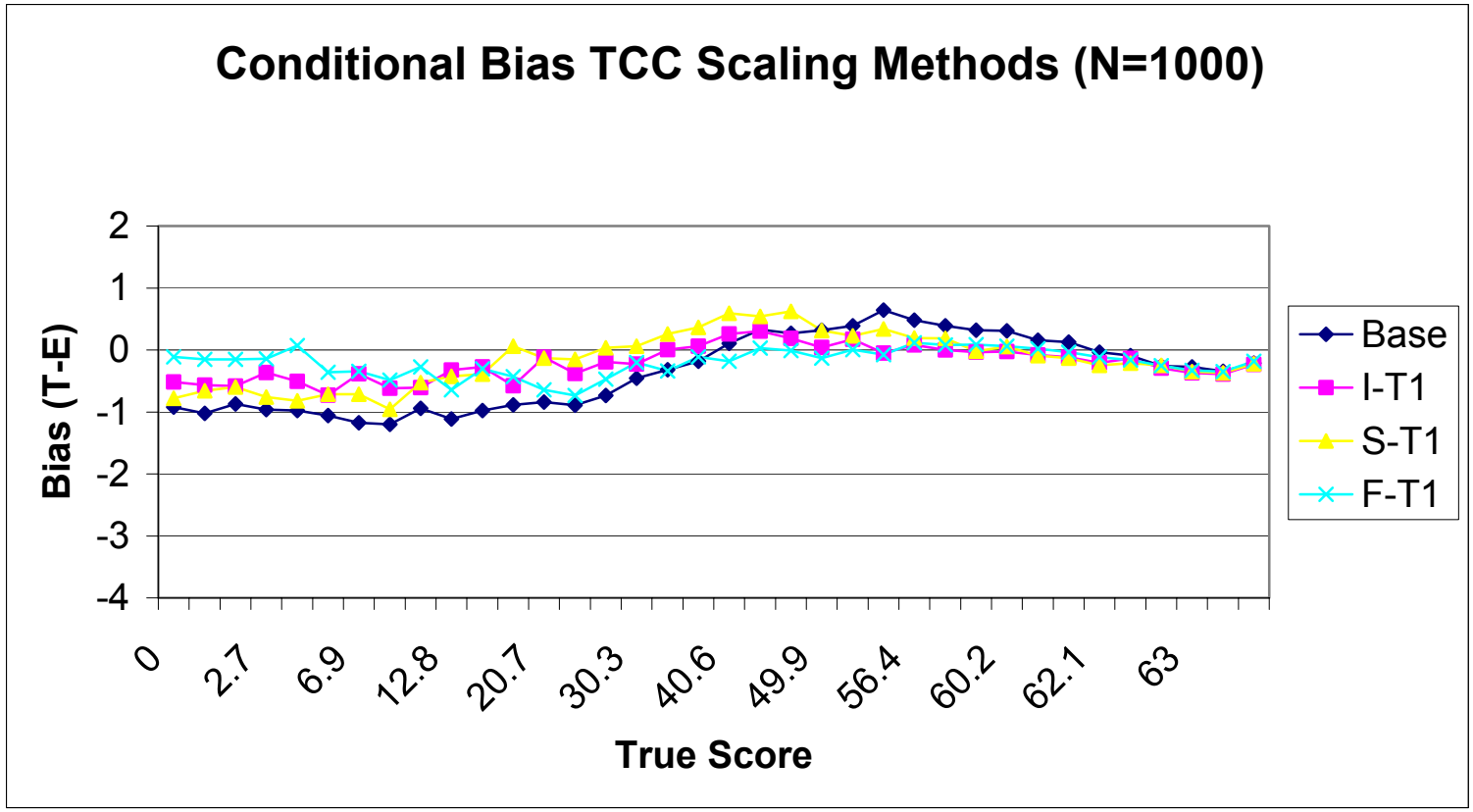

Figure 11
Conditional bias for TCC scaling methods (N=1000)

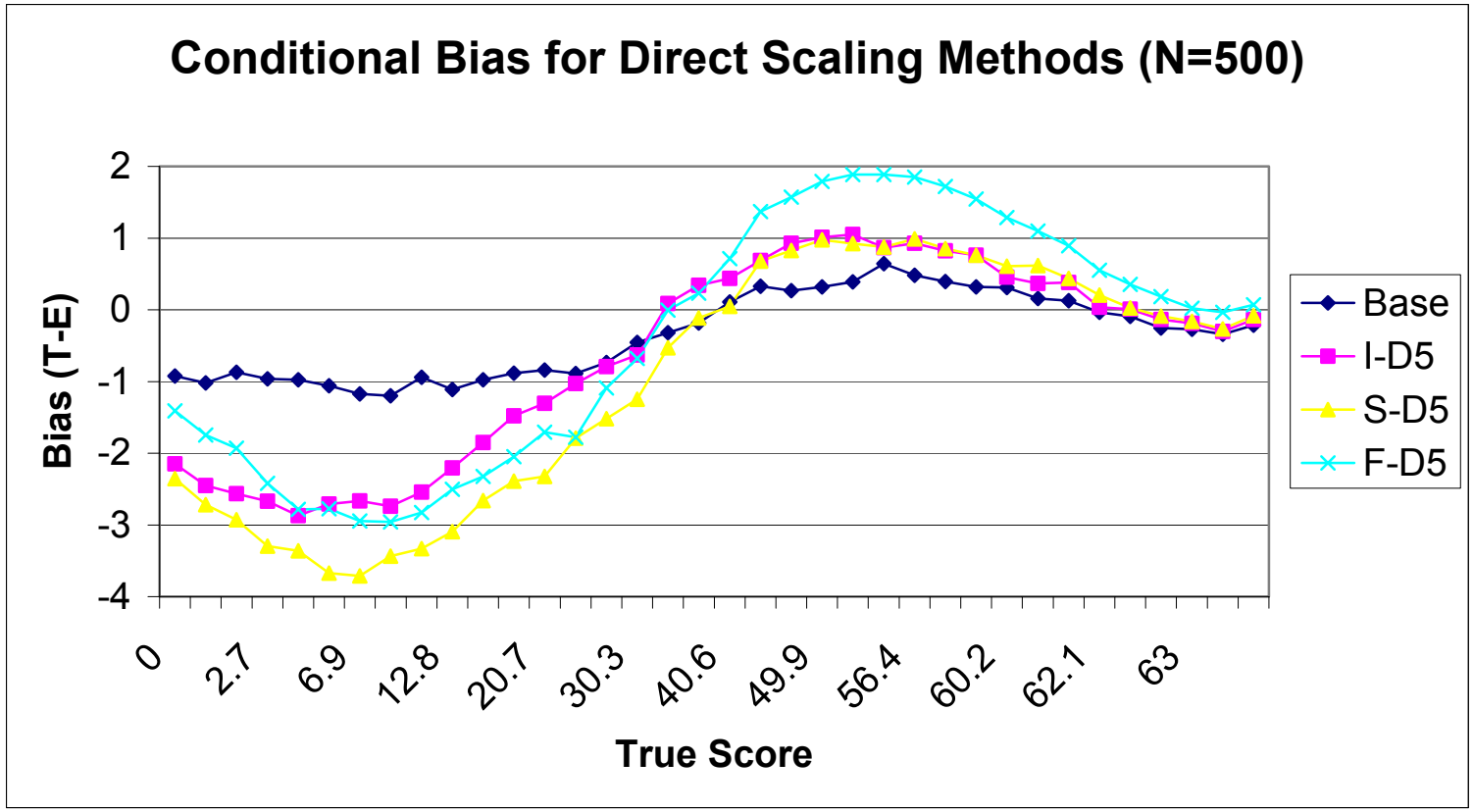

Figure 12
Conditional bias for direct scaling methods (N=500)

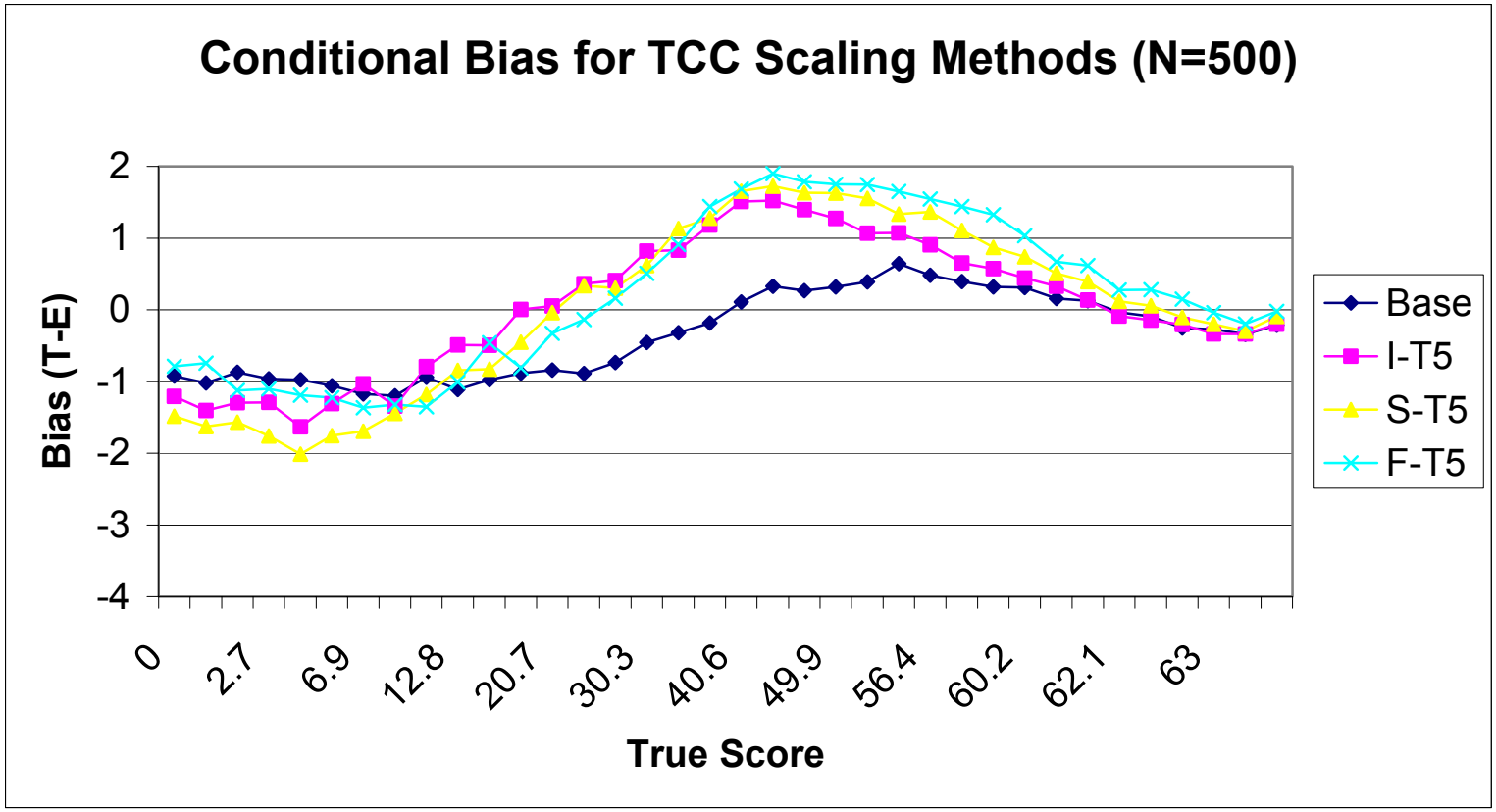

Figure 13
Conditional bias for direct scaling methods (N=1000)