# Updated Item Parameter Estimates Using Sparse CAT Data

Robert L. Smith
Saba Rizavi
Roxanna Paez
Ourania Rotou

Educational Testing Service

Paper presented at the annual meeting of the National Council on Measurement in Education
New Orleans, April 2002

# Updated Item Parameter Estimates Using Sparse CAT Data[1]

A computerized adaptive test (CAT) strongly depends on items being well modeled by the probability (IRT) model that underlies item selection and test scoring. The testing process will be less efficient and test scores both less precise and more biased to the extent that the IRT model fails to fit. Unfortunately, model parameters are usually calibrated from data collected under less than ideal circumstances. For example, some calibration data is collected from paper-and-pencil administrations despite the items' intended use on computer. Even when data is collected on computer, the calibration data is often collected in a linear administration although the items will be administered adaptively. Context effects (Kingston & Dorans, 1984) can undermine the validity of scores when items perform differently under pretest and operational conditions.

The ideal calibration data set would have items administered to large examinee samples under realistic testing conditions. Better item parameter estimates might be obtained if information from the adaptive administration could be incorporated into each item's calibration. However, while these conditions of delivery would help minimize context effects and reduce random error in the parameter estimates, CAT data may also introduce bias into the parameter estimates.

How best to incorporate CAT response data into the parameter estimates is not clear. A Bayesian IRT approach is one possibility for allowing CAT data to strengthen existing pretest parameter estimates. This may provide a mechanism for incorporating information about CAT item variation into the parameter estimates.

The goal in this study was to investigate whether augmenting the calibration of items using CAT data matrices produced estimates that were unbiased and improved the stability of the existing item parameter estimates.

## Method

### Design

Item parameter estimates from four pools of items constructed for operational use were used in the study. After allowing for some redundancy across pools, the final number of unique items was 1,392. All items had been previously calibrated using a three-parameter logistic model (3PL) from either paper-administration data or by seeding pretest items in a linear manner (i.e., first item first, second item second, etc.) within a CAT. Original calibration sample sizes varied from 600 to 1,500 test takers. All items had been calibrated using PARSCALE (Muraki & Bock, 1995). As a result, item priors were available for each item. The item prior in PARSCALE is composed of the parameter estimates, the means of the priors, and a variance-covariance matrix based on the slope, guessing and intercept parameters. The variance covariance matrix was

[1] The authors would like to thank Charlie Lewis and Tim Davey for many fruitful discussions and Kathy Sheehan for providing the specifics for the intercept-to-threshold transformation in PARSCALE.

transformed to include the slope, guessing and threshold prior to "true" item parameter generation.

Fifty sets of "true" parameter estimates were generated from the base item prior information. Each "true" set served as the parameter estimates for a CAT simulation that incorporated content constraints (Stocking & Swanson, 1993) and exposure control (Stocking & Lewis, 1998). The scored data matrices from each set of simulations (see below) were assembled into a sparse matrix that was then input to PARSCALE. Items were calibrated using the original (base) item priors. Fifty calibrations were performed, one for each set of simulations. The item (re)calibration produced posterior estimates for each set of parameters. The item characteristic curves (ICCs) and the parameters themselves were compared to the "true" estimates to determine whether root mean squared errors (RMSE) were reduced and whether any detectable bias was introduced into the parameter estimates.

Covariance transformation.

As noted above, the item parameters had been calibrated using PARSCALE. The item parameter covariance matrix for an item produced by PARSCALE is based on the slope, asymptote, and the intercept rather than the threshold parameter. Prior to parameter generation, the covariances including the intercept were transformed to the threshold (b) scale using the following procedure.

Let $\mathbf{X}$=[intercept, a, c] be the vector of item parameter estimates and let $\Sigma_{XX}$ be the posterior covariance matrix for an item. The elements of $\Sigma_{XX}$ are denoted as indicated below,

$$\Sigma_{XX} = \begin{bmatrix} s_{ii} s_{ia} s_{ic} \\ s_{ia} s_{aa} s_{ac} \\ s_{ic} s_{ac} s_{cc} \end{bmatrix}$$

Let $\mathbf{Y}$=[b, a, c]. The transformation function, $g$, from $\mathbf{X}$ to $\mathbf{Y}$ is defined as:

$$g(\text{int}) = (-1/a)\ \text{int} = b$$

$$g(a) = a$$

$$g(c) = c$$

A transformed matrix, defined below, is obtained,

$$\Sigma_{YY} = \begin{bmatrix} s_{bb} s_{ba} s_{bc} \\ s_{ba} s_{aa} s_{ac} \\ s_{bc} s_{ac} s_{cc} \end{bmatrix},$$

by applying the delta method as follows:

$$\Sigma_{YY} = \text{var}(g(X) = A\Sigma_{XX}A'$$

where $A$ is the matrix of partial derivatives of g(**X**).

After simplification the transformed covariance matrix, $\Sigma_{YY}$, may be written as,

$$\Sigma_{YY} = \begin{bmatrix} \frac{s_{ii} + 2bs_{ia} + b^2 s_{aa}}{a^2} & \frac{-s_{ia} - bs_{aa}}{a} & \frac{-s_{ic} - bs_{ac}}{a} \\ \frac{-s_{ia} - bs_{aa}}{a} & s_{aa} & s_{ac} \\ \frac{-s_{ic} - bs_{ac}}{a} & s_{ac} & s_{cc} \end{bmatrix}$$

Data generation.

Fifty sets of item parameters were generated for each item from the transformed item priors. This was accomplished by first drawing a random vector $\mathbf{z_r}$ from a standard multivariate normal distribution with mean and variance $(\mathbf{0}, \mathbf{I}_i)$, where $\mathbf{I}_i$ is a square matrix with 1s in the diagonal and zeros in the off-diagonals. This uncorrelated vector of draws was then transformed to the scale of the item with appropriate item covariances using the variance-covariance matrix $\Sigma_{YY}$ and the vector of parameter means using the following:

$$\mathbf{d_r} = \mathbf{v_i} \cdot \mathbf{z_r} + \mu_i$$

where $\mathbf{v_i}$ is a Choleski factor from the decomposition of the transformed variance-covariance matrix $\Sigma_{YY} = \mathbf{v}_i\mathbf{v}_i'$

Simulations.

Item responses were simulated for each pool to be included in the subsequent item calibration. Four simulations (one for each pool) were run for each calibration. Each set of four simulations used one set of the 50 sets of generated item parameters. Simulations were based on the weighted-deviation model (Stocking & Swanson, 1993) simultaneously taking into account content, item exposure and statistical criteria. The multinomial version of the Sympson-Hetter approach (Stocking & Lewis, 1998) was used to control item exposure. Items were selected to maximize information at the provisional ability estimate given the other constraints of the model.

One-thousand test takers were simulated (simulees) at each of 41 points[2] on the ability scale between ± 4.0 for a total of 41,000 simulated test takers. Prior to calibration this uniform distribution was sampled down to approximate the distribution of ability in the test-taking population. All 1,000 simulees were sampled at the highest point in the ability distribution; other abilities were proportionally less. This yielded 12,960 simulees per pool or 51,840 across the four pools for a given calibration.

Item calibrations.

The item responses from the sampled down simulations from the four pools were assembled in a sparse matrix as input for item calibration. A separate calibration was run for each of the 50 sets of generated parameters using PARSCALE. The original (base) item priors were used as priors for the item calibrations.

Evaluation.

Item parameter estimates were available from three sources: the base (original) parameters on which the generating priors are based; the generated parameter estimates (50 sets); and the posterior parameter estimates following the calibration of the simulated CATs (50 sets). Three comparisons are possible, the original item parameters (or ICCs) vs. the generated item parameters (or ICCs), the generated item parameters (or ICCs) vs. the calibrated item parameters (or ICCs), and the original item parameters (or ICCs) vs. the calibrated item parameters (or ICCs).

The primary interest is in the first two comparisons (generated-original, generated calibrated). The comparison of the original item parameters (or ICCs) with the generated parameters provides a measure of the expected variation in the parameter estimates prior to the inclusion of CAT information. This serves as a baseline unencumbered by CAT information. The second comparison of interest is between the generated (true) and the calibrated item parameters (ICCs). This comparison gives an assessment of the influence of incorporating CAT data into the item parameter estimates. However, there are a number of intervening factors that also come into play in this comparison.

Since CAT items are not delivered uniformly, sample size will have an influence on how (and why) the item parameters are changed. Items that are administered large numbers of times will have likelihoods that overpower the priors. In these cases, the influence of the item prior will be small. For items that are administered a relatively small number of times, the likelihood for an item will be overpowered by the prior for that item. In the most extreme case where the item is not administered to anyone, the parameter estimates for the calibrated item will equal the item parameters (and ICC) of the prior. This comparison, thus, reduces to the comparison between the generated and the original parameter estimates in this case.

[2] All operational items had been calibrated based on 41 quadrature points. Simulations were conducted at the same number of quadrature points so that there was agreement between the number of abilities in the simulations and in the calibrations.

Comparisons between ICCs were evaluated with a root mean squared difference (RMSD) between two ICCs. Item parameter comparisons were assessed using the root mean squared error (RMSE). In addition, bias analyses were conducted on the parameter estimates. The RMSD for the ICC for the generated-original comparison was defined as:

$$RMSD_{go} = \sqrt{\left[\int_{-4}^{4} P(\theta \mid a_{tr}, b_{tr}, c_{tr}) - P(\theta \mid a_o, b_o, c_o)]^2 \, d\theta\right]},$$

where *t* stands for the true parameter, *r* stands for the replication, *o* stands for the original parameter. The integral is approximated based on 41 quadrature points. Similarly, the RMSD for the ICC for the generated-calibrated comparison was defined as:

$$RMSD_{gc} = \sqrt{\left[\int_{-4}^{4} [P(\theta \mid a_{tr}, b_{tr}, c_{tr}) - P(\theta \mid a_{cr}, b_{cr}, c_{cr})]^2 \, d\theta\right]},$$

where *t* stands for the true parameter, *r* stands for the replication, *c* stands for the calibrated parameter.

Unweighted RMSE were computed[3]. In the generated-original comparison, the mean ICC for the generated curves was compared to the single original ICC from which they were generated.

The impact of CAT data on the a-, b-, and c-parameter estimates was assessed through root mean squared error (RMSE) and bias, where RMSE for the generated-original comparison is defined as,

$$RMSE\left(\hat{\beta}_i\right) = \sqrt{\frac{1}{50}\sum_{r=1}^{50}\left(\beta_{tr} - \hat{\beta}_o\right)^2},$$

where $\beta_i$ is the parameter for item *i* on replication *r*. The subscript *t* stands for the true (generated) parameter, while *o* stands for original parameter. The RMSE for the generated-calibrated comparison is defined as,

$$RMSE\left(\hat{\beta}_i\right) = \sqrt{\frac{1}{50}\sum_{r=1}^{50}\left(\beta_{tr} - \hat{\beta}_{cr}\right)^2},$$

where $\beta_i$ is the item parameter of item *i* on replication *r*. The subscript *t* stands for the true (generated) parameter, while *c* stands for calibrated estimate using CAT data. Similarly, the bias for the parameter estimates in the generated-original comparison is defined as,

[3] Both weighted and unweighted RMSDs were computed, however, only the unweighted RMSDs is reported. The weighted form of the RMSD tended to mask differences that were present. Even if these effects later proved to have little substantive impact, we wanted to know they were present.

$$Bias((\hat{\beta}_i) = \frac{1}{50}\sum_{r=1}^{50}(\beta_{tr} - \hat{\beta}_o),$$

and bias for the parameter estimates in the generated-calibrated comparison is defined as,

$$Bias((\hat{\beta}_i) = \frac{1}{50}\sum_{r=1}^{50}(\beta_{tr} - \hat{\beta}_{cr}).$$

## Results

Prior to examining the influence of CAT data on the parameter estimates or ICCs, it is important to establish that the generated parameters did not contain bias from the start. Figures 1-3 show the spread of estimates around the original (base) value by item difficulty. Figures 4-6 show the relationship between the original (base) and generated parameters. The b-parameter mean differences tend to be evenly distributed about zero with greater spread as item difficulty increases, but the b-parameters do not appear to be biased here. Similarly, the a-parameter differences tend to be evenly distributed about zero with greater spread for easier and more difficult items. The c-parameters also do not show any bias, though there is less spread in the c-parameters for more difficult items. This is supported by the means bias for the a-, b-, and c-parameters found in Table 1 which is very close to zero. Finally, there is good correspondence between the mean of the 50 generated parameters and the original parameters as shown in the diagonal plots (Figures 4-6).

Comparing the original estimates to the generated values gives a measure of the expected variation in the estimates. Figure 7 shows the variation in ICCs as a function of item difficulty. Items with difficulties between $\pm 2$ tend to be less variable than easier or more difficult items, with the greatest variation found for the most difficult items. Figure 8 shows a similar pattern for the b-parameters. B-parameters of middle difficulty ($\pm 2$) also tend to be less variable than easier or more difficult items. For the a-parameters, again, more difficult items tend to have larger RMSEs than easier items (Figure 9). The c-parameters of more difficult items appear better estimated than the c-parameters for easier items (Figure 10). These patterns are common for standard calibration designs where pretests are randomly assigned to test takers. These plots serve as a baseline against which to compare item parameter estimates that have incorporated CAT variation.

Our primary comparison of interest is between the true (generated) estimates and the estimates after CAT data has been incorporated (post calibration estimates). These comparisons should show what influence, if any, CAT data has on the parameter estimates or the ICC as a whole. Table 1 presents average RMSEs between the original and the mean of the generated parameters and between the mean-generated and mean-calibrated parameters (or ICCs). The RMSDs for the ICCs appear to be smaller for the generated-calibrated comparison which contains CAT information, than the baseline generated-original comparison, suggesting overall error is reduced with the inclusion of CAT information (compare Figure 11 with Figure 7). At the parameter level, there is a sizeable

decrease in the RMSE for the b-parameter suggesting that inclusion of the CAT data has "improved" the b-parameter estimates (see Figure 12, compare with Figure 8). The RMSEs for the a-parameters also appear to be slightly smaller when CAT data is incorporated into the parameter estimates.

The improved precision appears to be at the expense of some increase in the bias for both the a- and b-parameter estimates, however. Figure 13 shows the bias in the generated versus calibrated b-parameter estimates. Both the easiest and most difficult items tend to be downwardly biased (compare with Figure 2 for the generated versus original differences). Figure 14 shows that the bias in the a-parameter estimates appears related to the difficulty of the item (compare with Figure 1). Items with difficulties below 0 do not appear to be biased. Items from 0 to 2.0 tend to be downwardly biased, while those above 2.0 tend to be upwardly biased. Figure 15 shows the bias in the c-parameter. These too, are somewhat downwardly biased particularly for items of middle difficulty (compare with Figure 3).

Sample size, RMSE and bias.

As mentioned above, the number of exposures that an item receives will affect the relationship between the prior and the likelihood of a calibrated item. Consequently, there was interest in examining whether there was a relationship between sample size and RMSE or bias. Presumably, if CAT data introduces bias, larger amounts of CAT data should introduce larger amounts of bias. Similarly, if CAT data reduces error in the estimates, then larger amounts of CAT data should reduce the RMSE. Table 2 shows the correlation between the average calibration sample size for an item with the average RMSD for the ICCs, and with the RMSE and average bias for each of the parameters. Moderate negative relationships are observed between sample size and RMSD for the ICCs (-.47), and between the RMSE for the b-parameters (-.35) and c-parameters (-.21), suggesting that error in the estimates is reduced with increased sample size for the ICCs and the b- and c-parameter estimates. The plot of the RMSD for the ICCs for the generated-calibrated comparison also appears to show this reduction across all sample sizes when compared with the (generated-original) base RMSE plot (see Figures 16 and 17, respectively) as do the plots for the b-parameter estimate (Figures 18 and 19). The plots for the c-parameter show larger RMSEs following the incorporation of CAT data into the estimates, but it does not appear related to the number of exposures (Figures 20 and 21). The RMSE for the a-parameters also appears unaffected by sample size (Figures 22 and 23).

Figures 24-26 graphically display the relationship between mean bias and average sample size for the item parameters for an item. While it appears that mean bias decreases with increased sample size for the b-parameter, mean bias appears to increase with increased sample size for the a- and c-parameter estimates.

There is one caveat about the impact of sample size that suggests it should be interpreted cautiously. In a CAT there is a relationship between items that receive high

exposures (even with exposure control) and item difficulty. Items of middle to high difficulty tend to receive greater exposure than items that are below middle difficulty as shown in Figure 27. Thus, increased precision or reduced bias may be more a function of an item's difficulty than the number of times it is administered.

## Discussion

The present study sought to examine the influence of CAT data on IRT 3PL item characteristic curves overall and on item parameter estimates for the 3PL model in a Bayesian context. The error in the ICC as a whole and in the estimates of item difficulty appears to be reduced when CAT data information is incorporated into the estimation of this parameter. However, bias also seems to be increased for the item difficulty parameter, particularly for the easiest and most difficult items.

CAT data also appears to introduce bias into the slope parameter. However, the amount and direction of the bias appears related to item difficulty. Easy items (item difficulties below 0) do not appear to be biased, while those with difficulties between 0 and 2.0 appear downwardly biased and those with difficulties above 2.0 appear upwardly biased. The incorporation of CAT information into the a-parameter estimate also seems to decrease the precision in the slope parameter estimate.

There is some evidence that error is reduced in the threshold (b-parameter) and asymptote (c-parameter) estimates as an item receives more exposure. Also, bias in the b-parameter estimates appears to decrease as an item receives more exposure, while it seems to increase for the a-parameter and c-parameter estimates. However, one must be cautious about attributing this merely to the number of exposures because there is a relationship between the "number of hits" an item receives and item difficulty. Since items of middle to high difficulty receive more exposures than extremely easy items, and it may be characteristics related to difficulty rather than number of exposures that causes this to happen. As seen earlier, there is less bias for the middle difficulty items than for either the extremely easy or the extremely difficult items.

These results suggest that the error in estimating ICCs may be reduced by incorporating CAT information into the calibration process. It appears that much of this improvement can be attributed to the increased precision in the estimation of the b-parameter. There does not appear to be much improvement in the estimation of the slope and asymptote parameters. The b-parameters appear to show some bias for the easy and more difficult items. Bias in the a-parameters appears related to item difficulty with difficult items biased downward and extremely difficult items biased upward. The c-parameters also appear to be biased downward, particularly for items of middle difficulty. Given this, perhaps the inclusion of CAT data might prove most fruitful when the Rasch model is used. In the future, however, the issue of whether the influence of CAT data on item parameter estimates translates into real score differences needs to be further examined.

## References

Kingston, N. M. & Dorans, N. J. (1984). Item location effects and their implications for IRT equating and adaptive testing. *Applied Psychological Measurement*, *8*, 147-154.

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

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

Table 1
RMSE and bias statistics by comparison group and estimate

| | Generated -Original | | | | Generated -Calibrated | | | |
|---|---|---|---|---|---|---|---|---|
| | RMSE | | Bias | | RMSE | | Bias | |
| Estimate | Mean | Std Dev | Mean | Std Dev | Mean | Std Dev | Mean | Std Dev |
| ICC | .035 | .011 | - | - | .027 | .008 | - | - |
| a | .120 | .057 | .000 | .019 | .115 | .048 | -.019 | .031 |
| b | .173 | .109 | .000 | .028 | .121 | .090 | -.016 | .036 |
| c | .037 | .012 | .000 | .005 | .039 | .014 | -.005 | .007 |

Table 2
Correlation between RMSD (ICC), RMSE, and average bias (parameters) with average calibration sample size (generated-calibrated comparison)

| Estimate | RMSD/RMSE | Bias |
|---|---|---|
| ICC | -.47 | - |
| a | .01 | -.43 |
| b | -.35 | .22 |
| c | -.21 | -.24 |

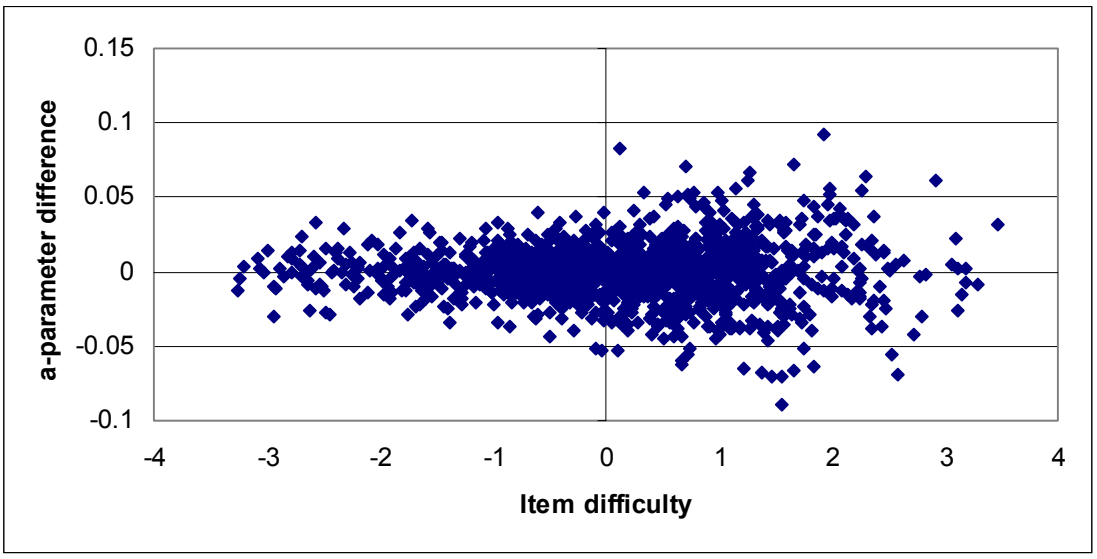


Figure 1. Scatterplot showing the relationship between a-parameter bias (generated – original) and item difficulty.

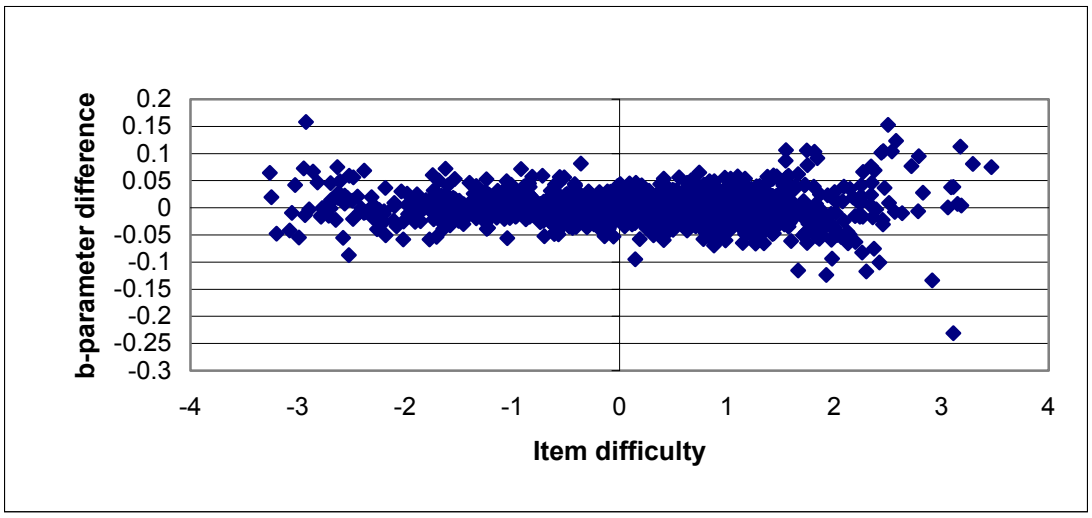


Figure 2. Scatterplot showing the relationship between b-parameter bias (generated – original) and item difficulty.

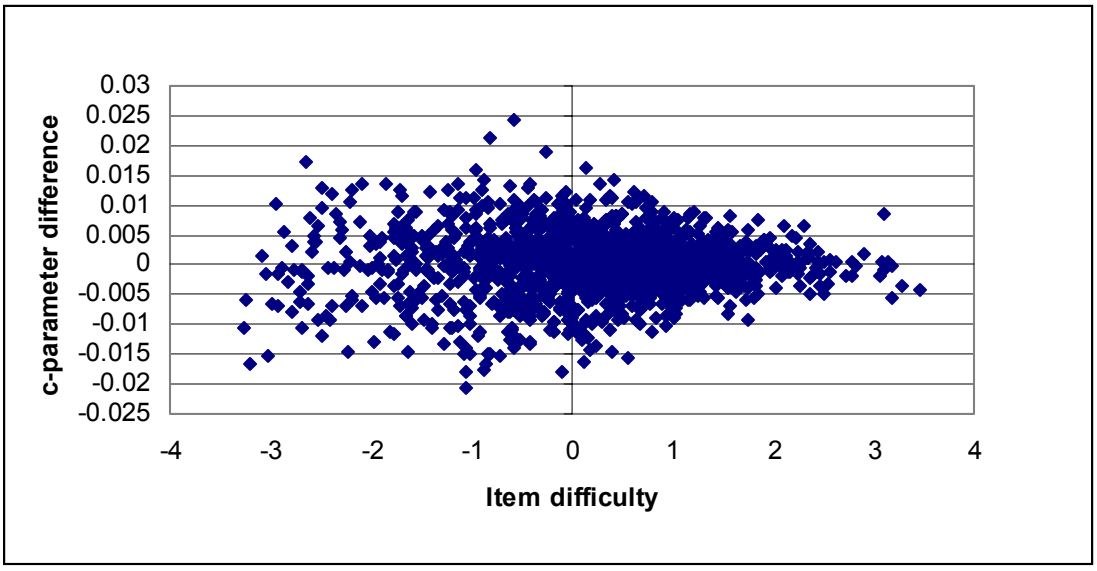


Figure 3. Scatterplot showing the relationship between c-parameter bias (generated – original) and item difficulty.

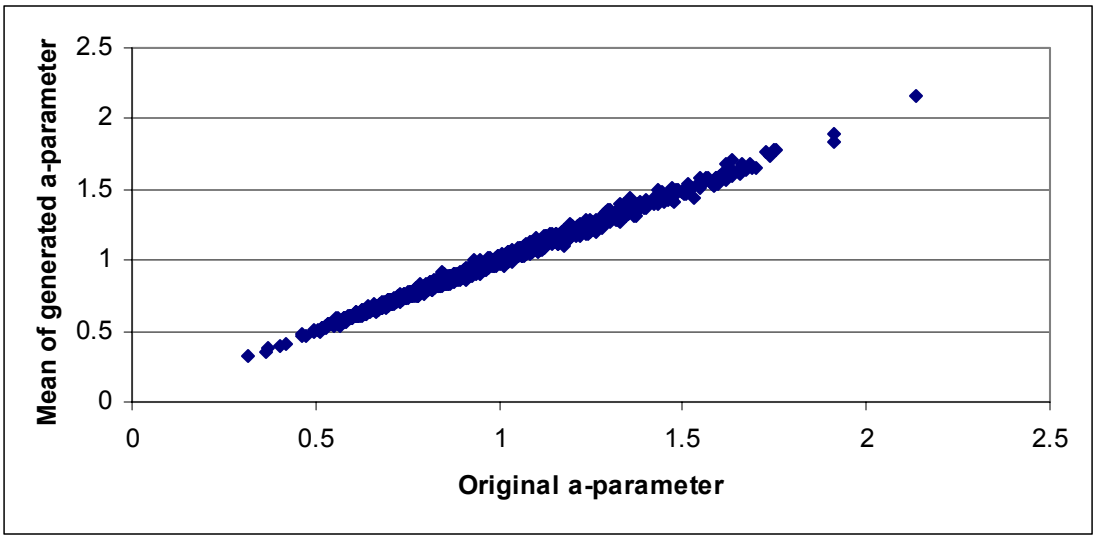


Figure 4. Scatterplot showing the relationship between original a-parameter and the mean of the generated a-parameters.

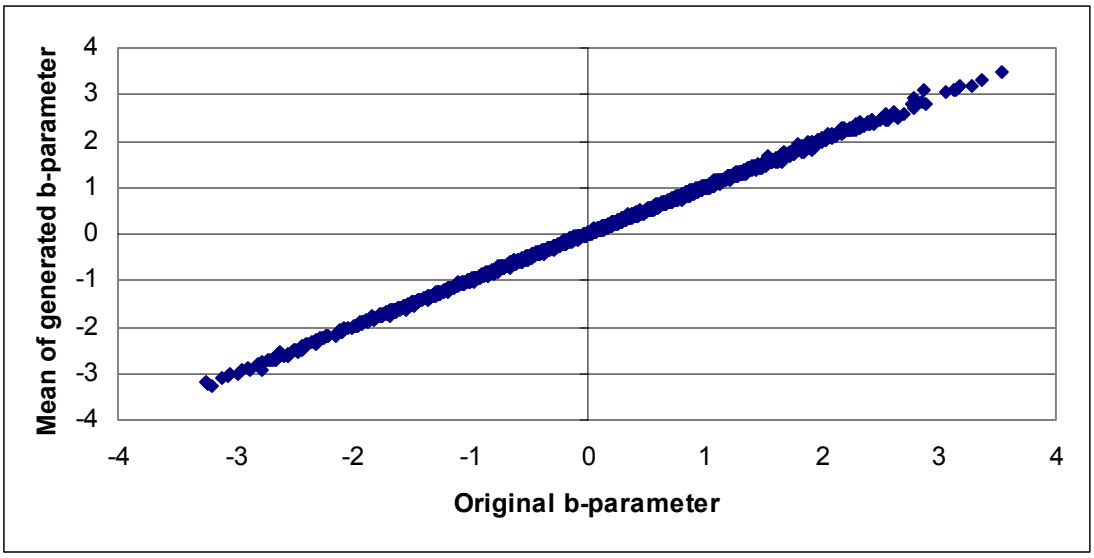


Figure 5. Scatterplot showing the relationship between original b-parameter and the mean of the generated b-parameters.

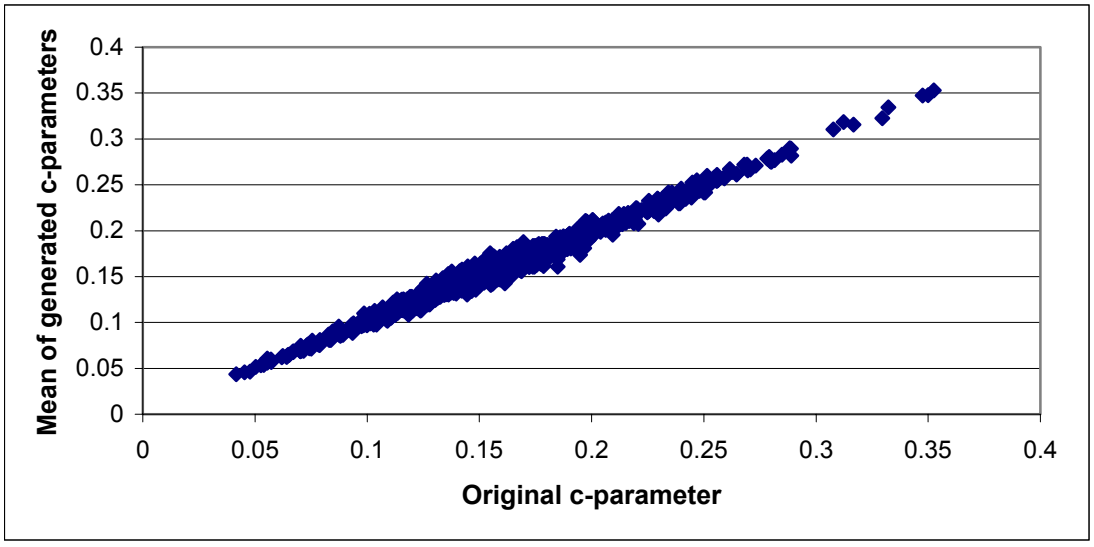


Figure 6. Scatterplot showing the relationship between original c-parameter and the mean of the generated c-parameters.

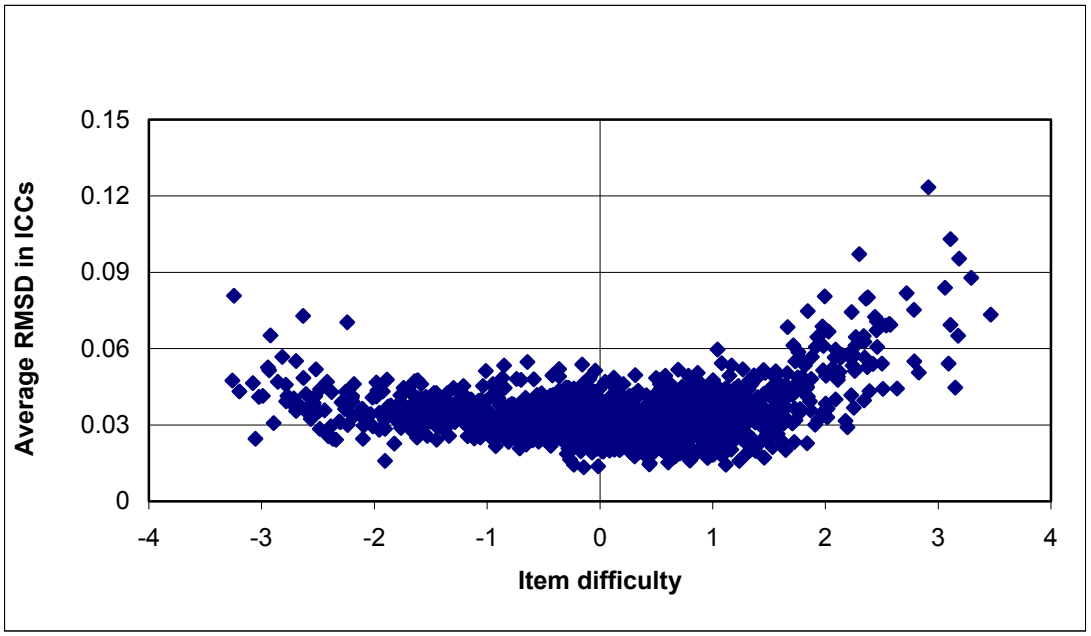


Figure 7. Scatterplot showing the relationship between the RMSD in ICCs for the b-parameter (generated – original) and item difficulty.

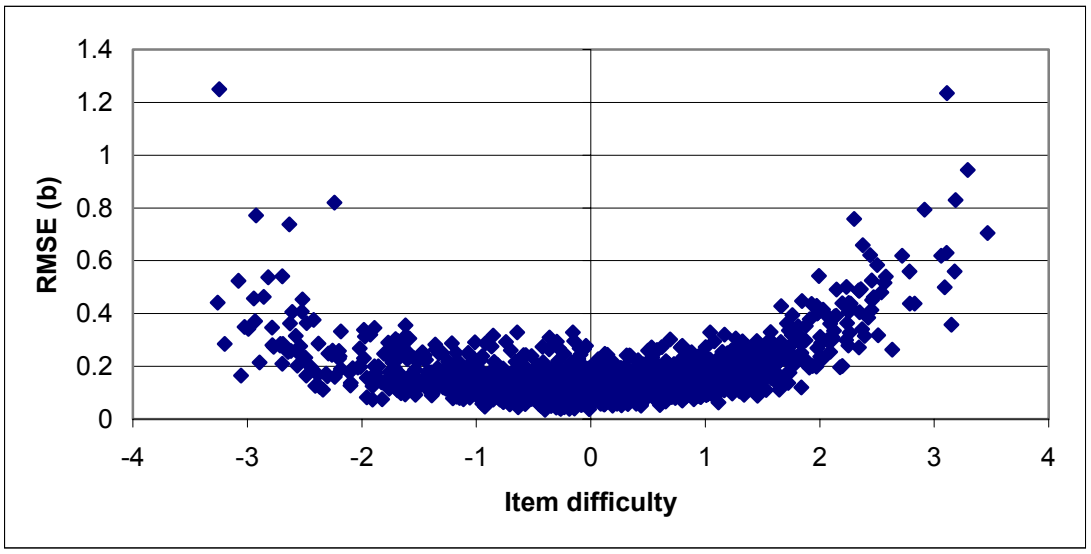


Figure 8. Scatterplot showing the relationship between the RMSE for the b-parameter (generated – original) and item difficulty.

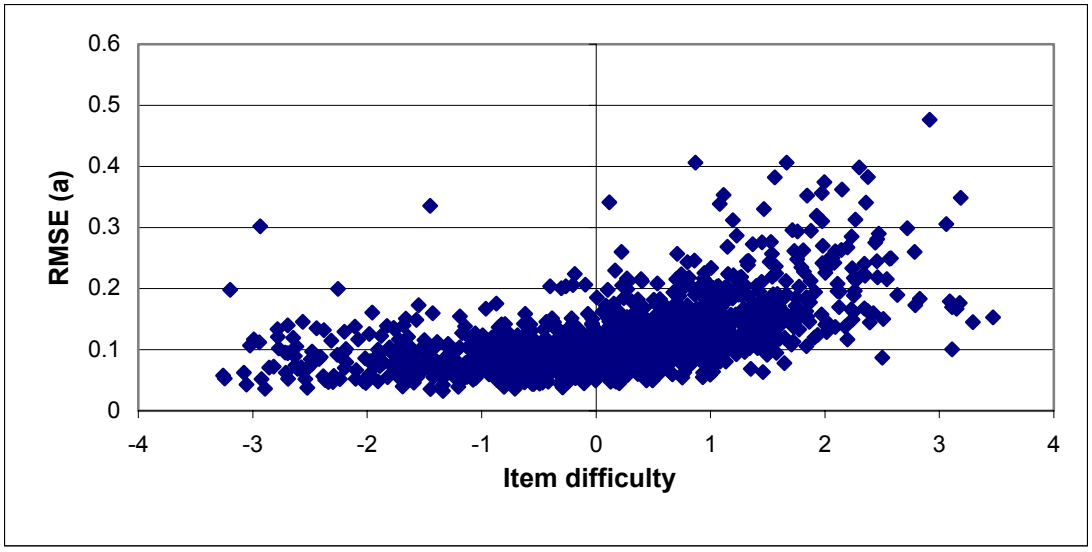


Figure 9. Scatterplot showing the relationship between the RMSE for the a-parameter (generated – original) and item difficulty.

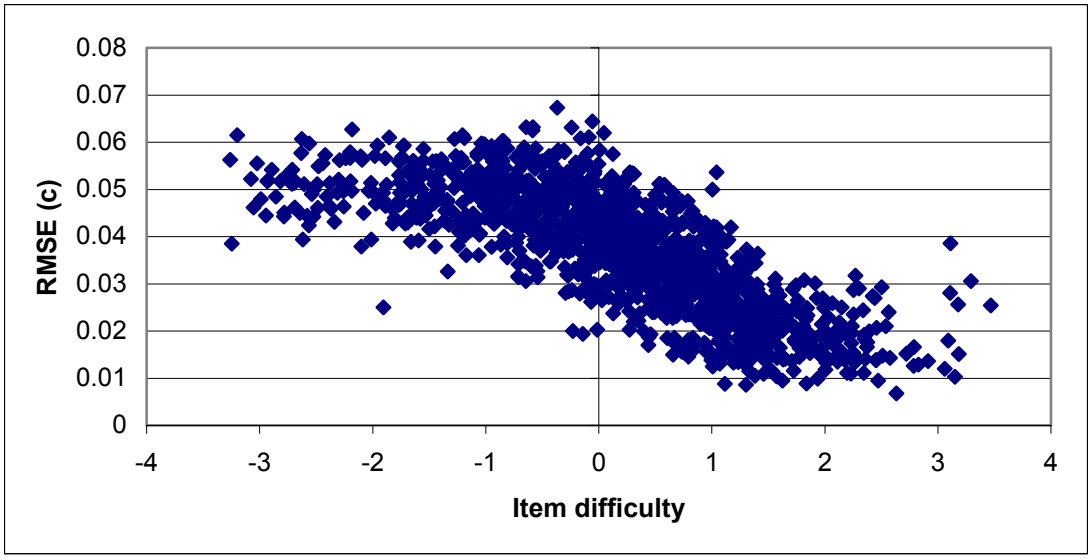


Figure 10. Scatterplot showing the relationship between the RMSE for the c-parameter (generated – original) and item difficulty.

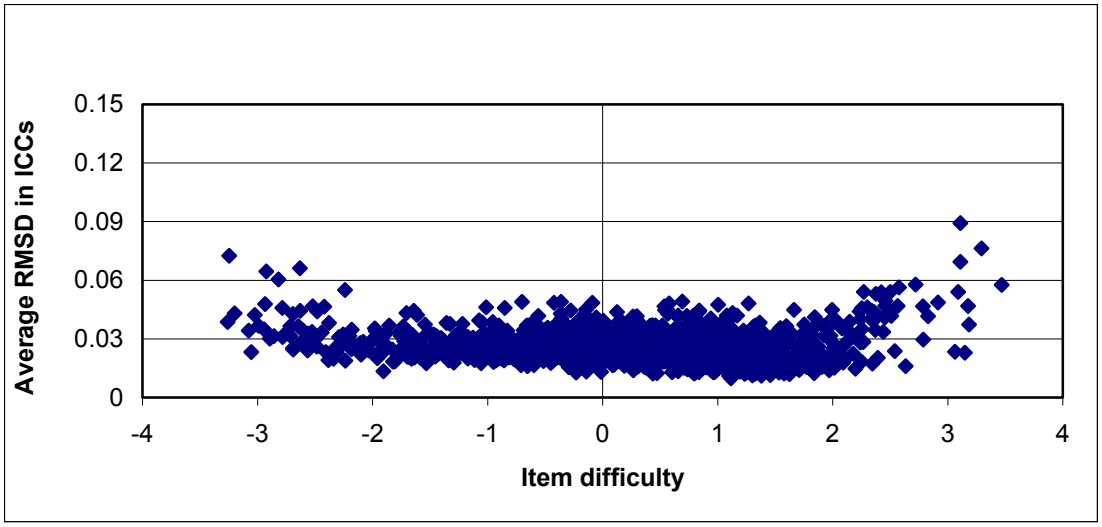


Figure 11. Scatterplot showing the relationship between the RMSD in ICCs (generated – calibrated) and item difficulty.

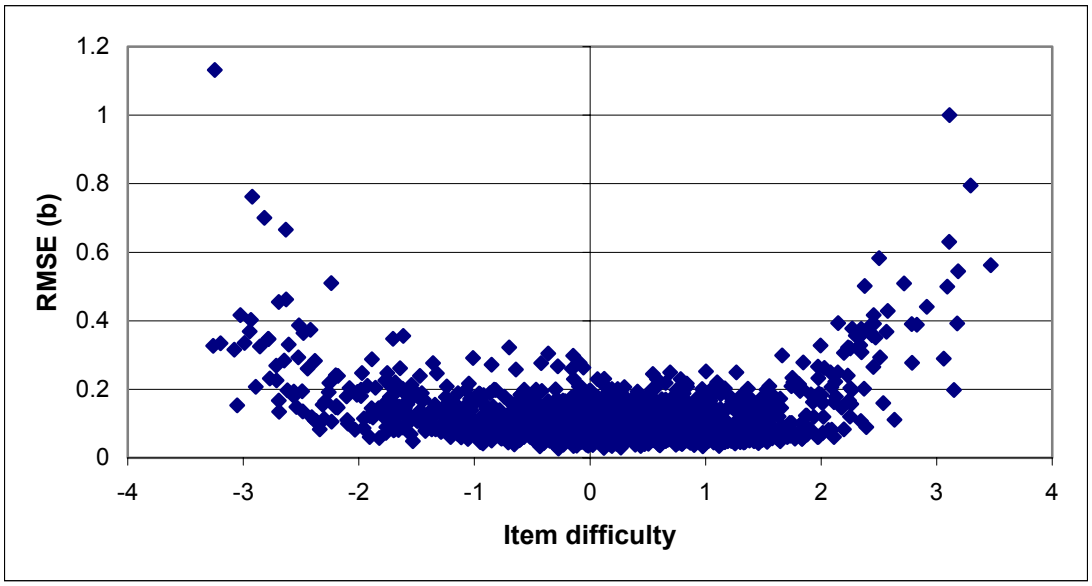


Figure 12. Scatterplot showing the relationship between the RMSE for the b-parameter (generated – calibrated) and item difficulty.

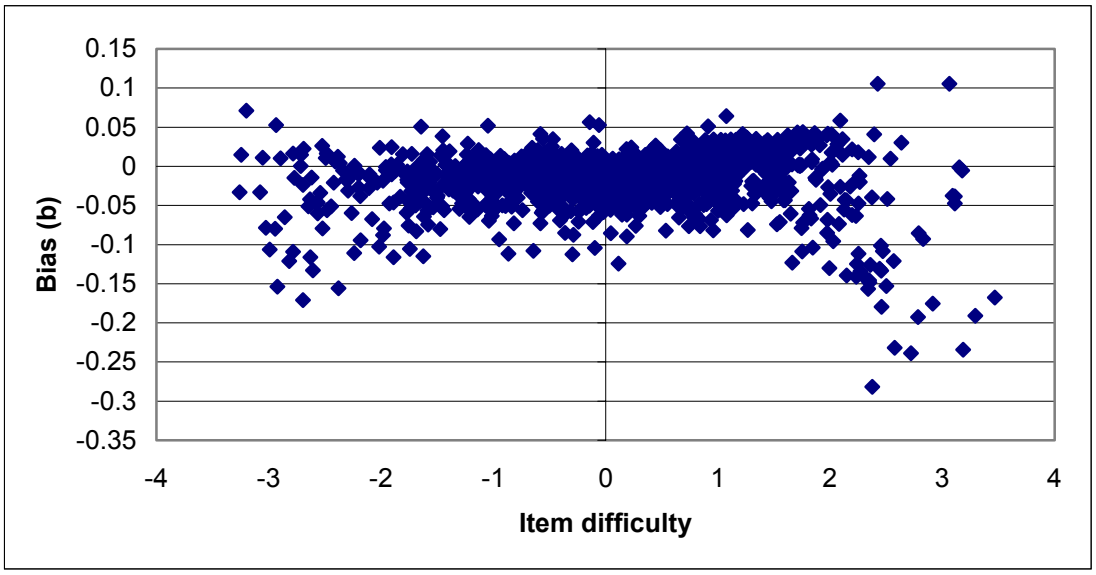


Figure 13. Scatterplot showing the relationship between b-parameter bias (generated – calibrated) and item difficulty.

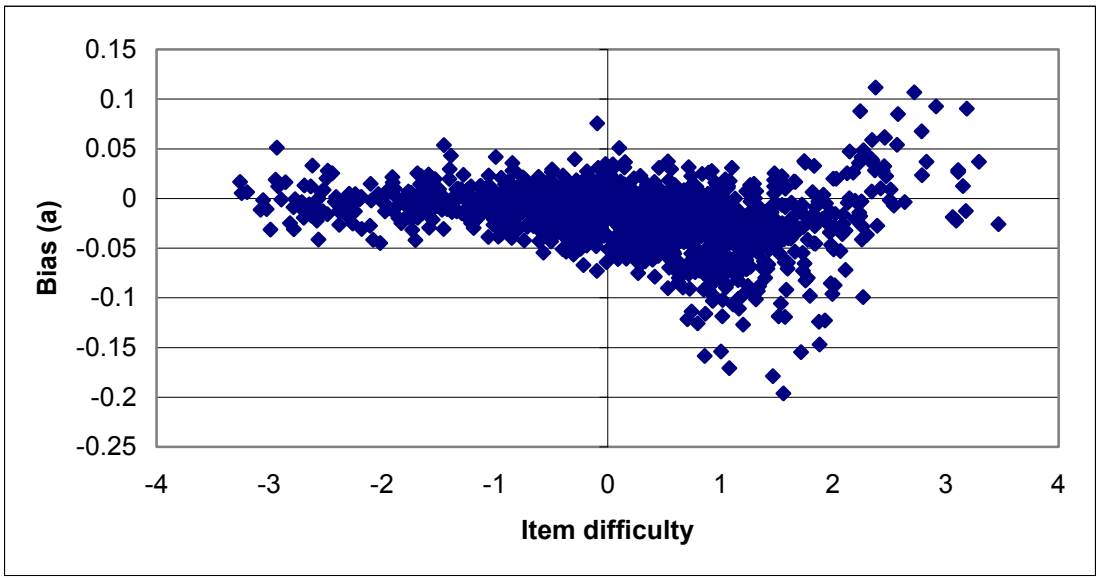


Figure 14. Scatterplot showing the relationship between a-parameter bias (generated – calibrated) and item difficulty.

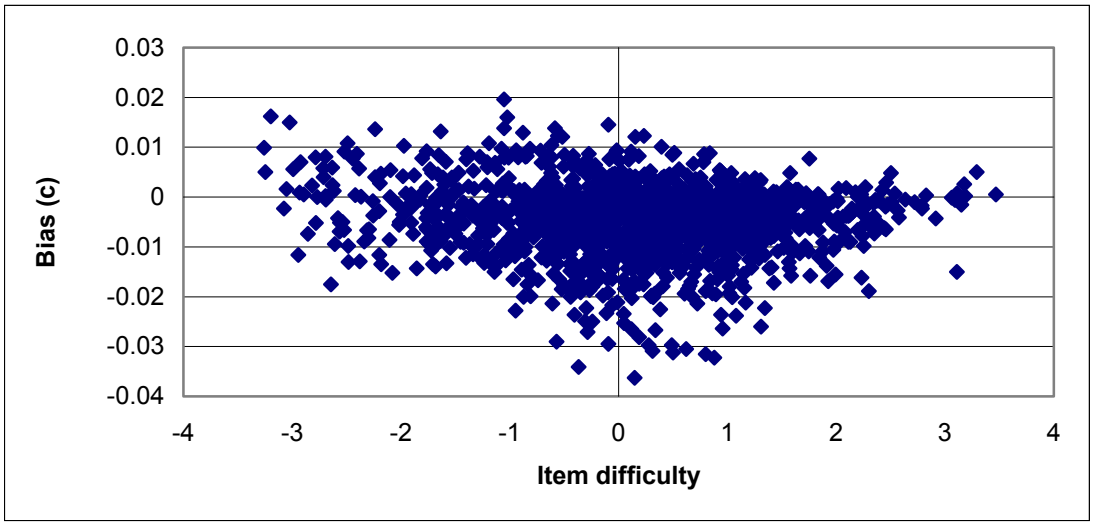


Figure 15. Scatterplot showing the relationship between c-parameter bias (generated – calibrated) and item difficulty.

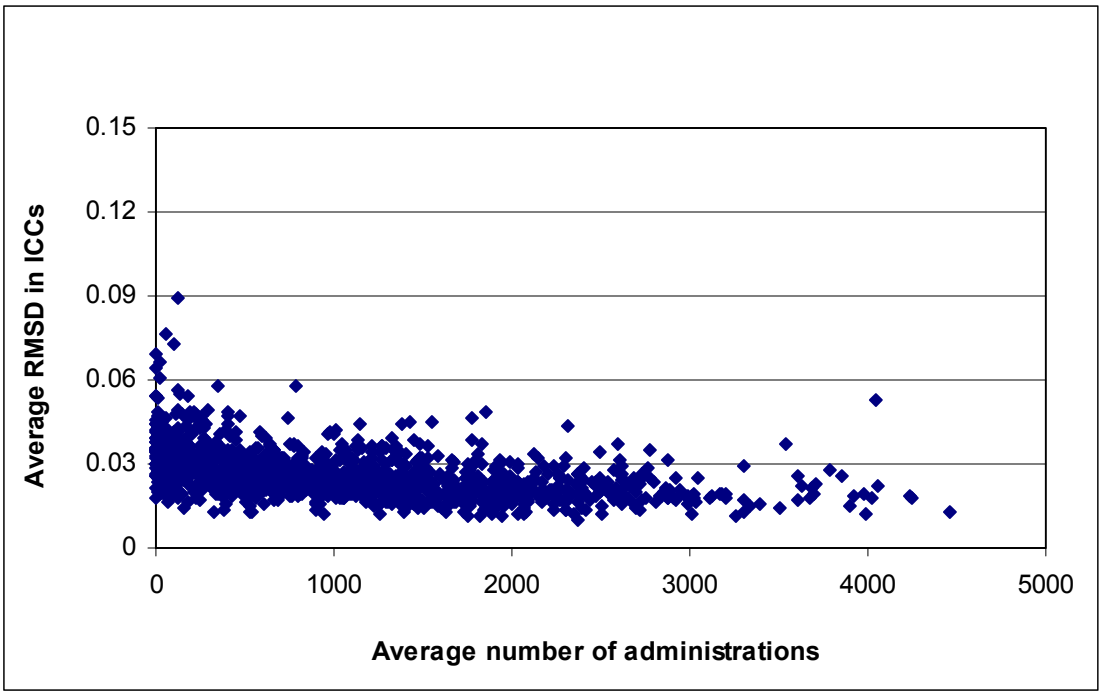


Figure 16. Scatterplot showing the relationship between the RMSD for the ICCs (generated – calibrated) and the average number of administrations for an item.

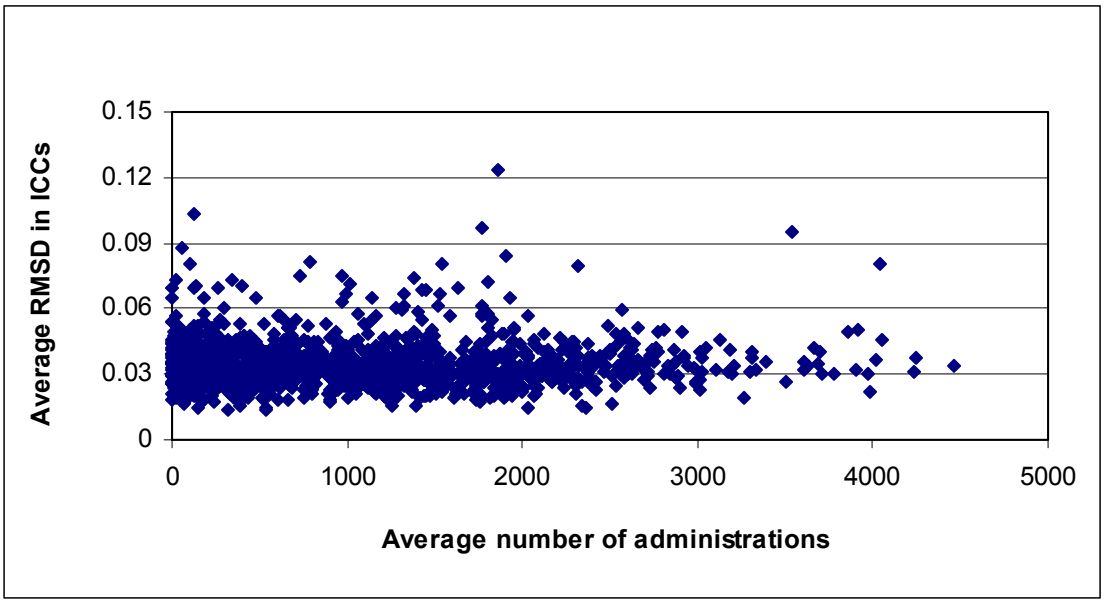


Figure 17. Scatterplot showing the relationship between the RMSD for the ICCs (generated – original) and the average number of administrations for an item.

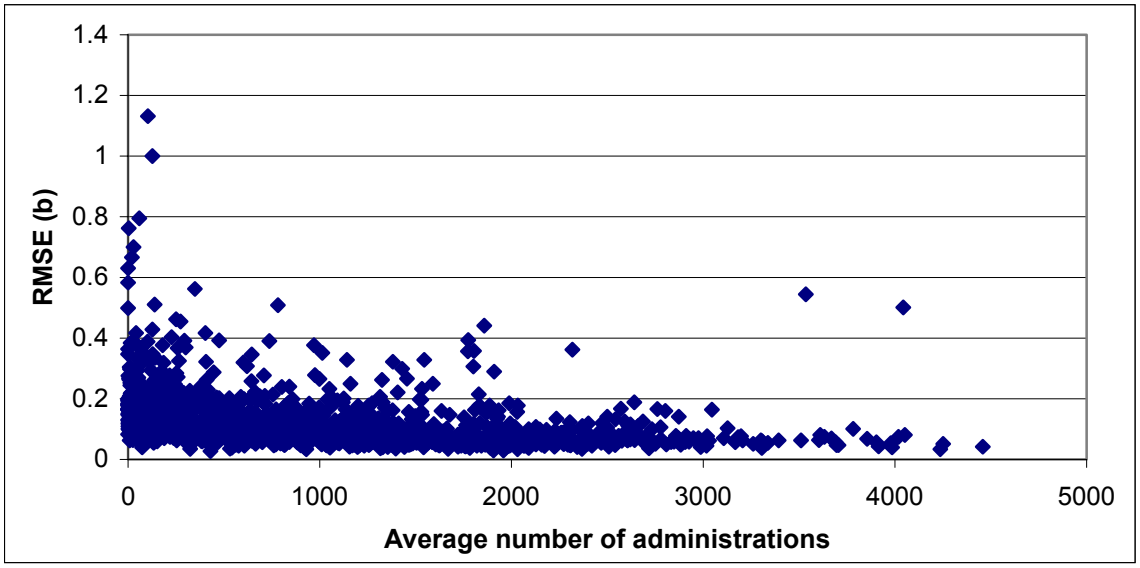


Figure 18. Scatterplot showing the relationship between the RMSE (generated – calibrated) for the b-parameter and the average number of administrations for an item.

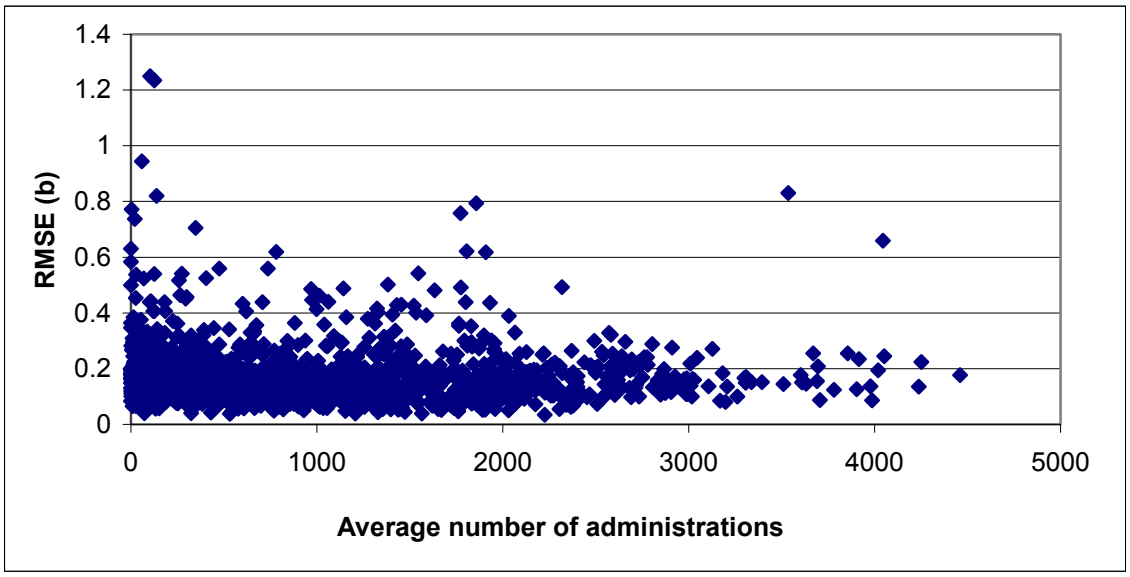


Figure 19. Scattergram showing the relationship between the RMSE (generated – original) for the b-parameter and the average number of administrations for an item.

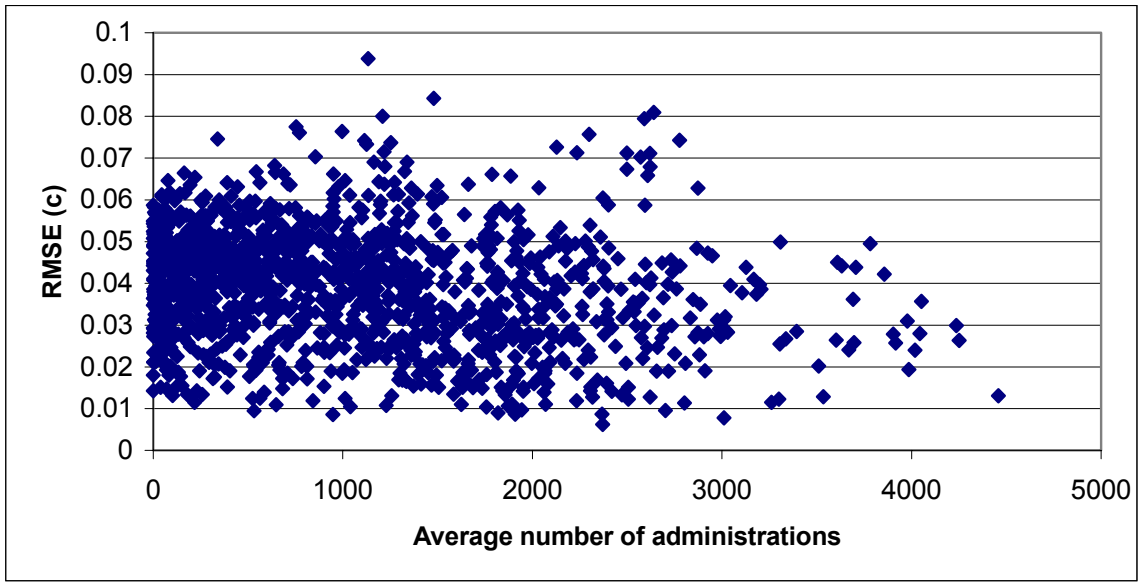


Figure 20. Scatterplot showing the relationship between the RMSE (generated – calibrated) for the c-parameter and the average number of administrations for an item.

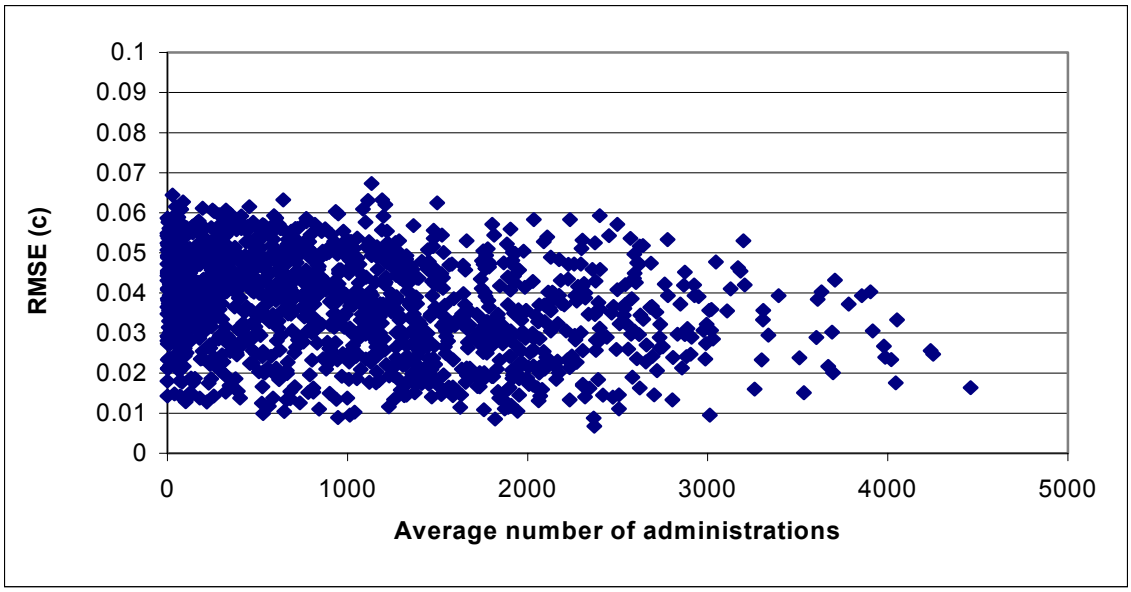


Figure 21. Scattergram showing the relationship between the RMSE (generated – original) for the c-parameter and the average number of administrations for an item.

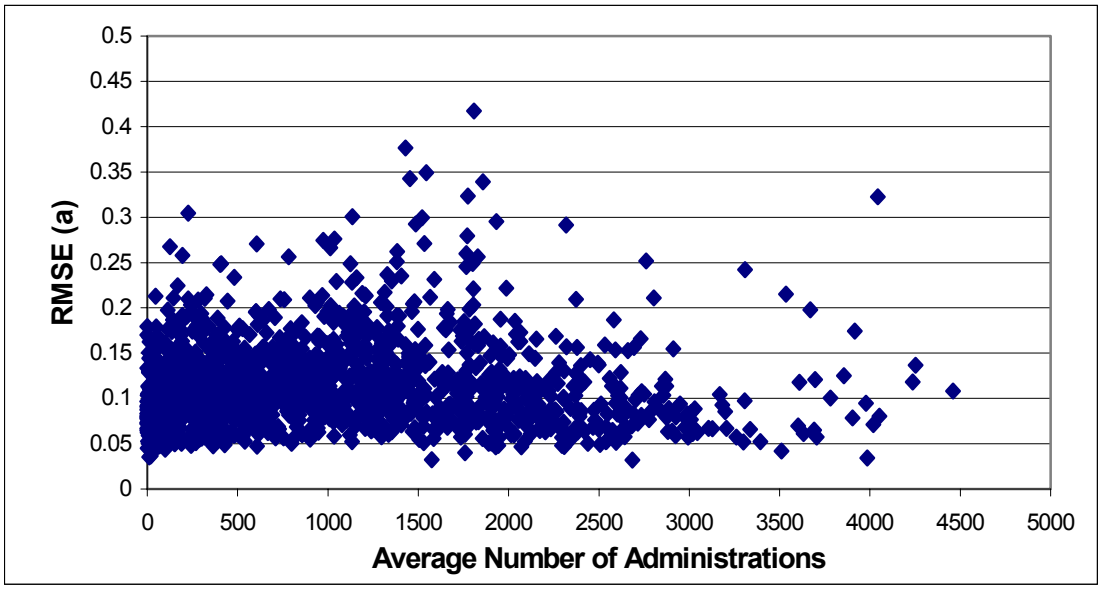


Figure 22. Scatterplot showing the relationship between the RMSE (generated-calibrated) for the a-parameter and the average number of administrations for an item.

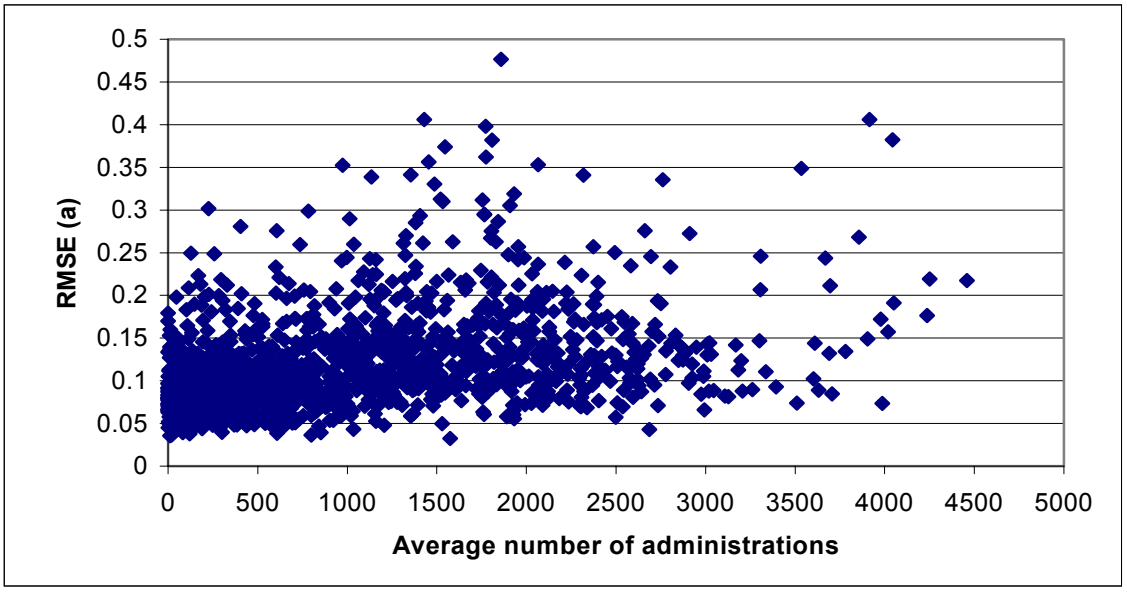


Figure 23. Scattergram showing the relationship between the RMSE (generated – original) for the a-parameter and the average number of administrations for an item.

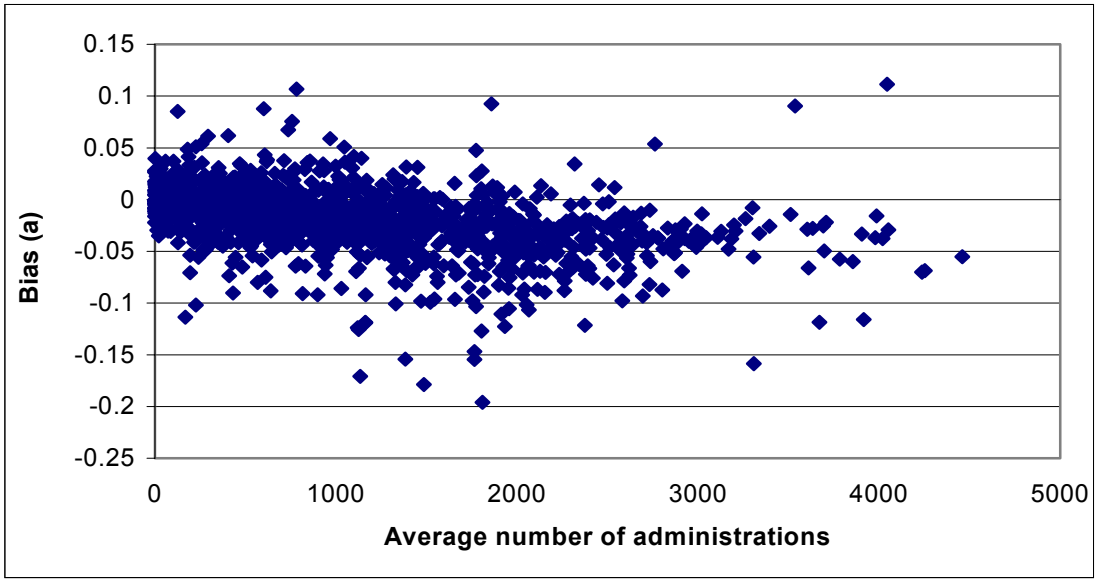


Figure 24. Scatterplot showing the relationship between the bias (generated –calibrated) for the a-parameter and the average number of administrations for an item.

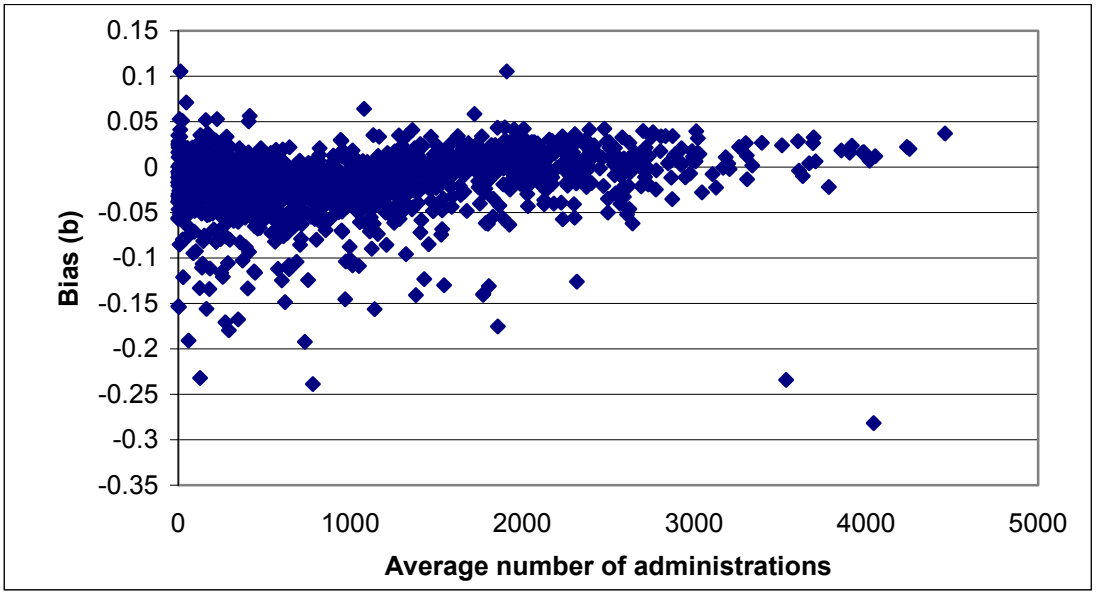


Figure 25. Scatterplot showing the relationship between the bias (generated –calibrated) for the b-parameter and the average number of administrations for an item.

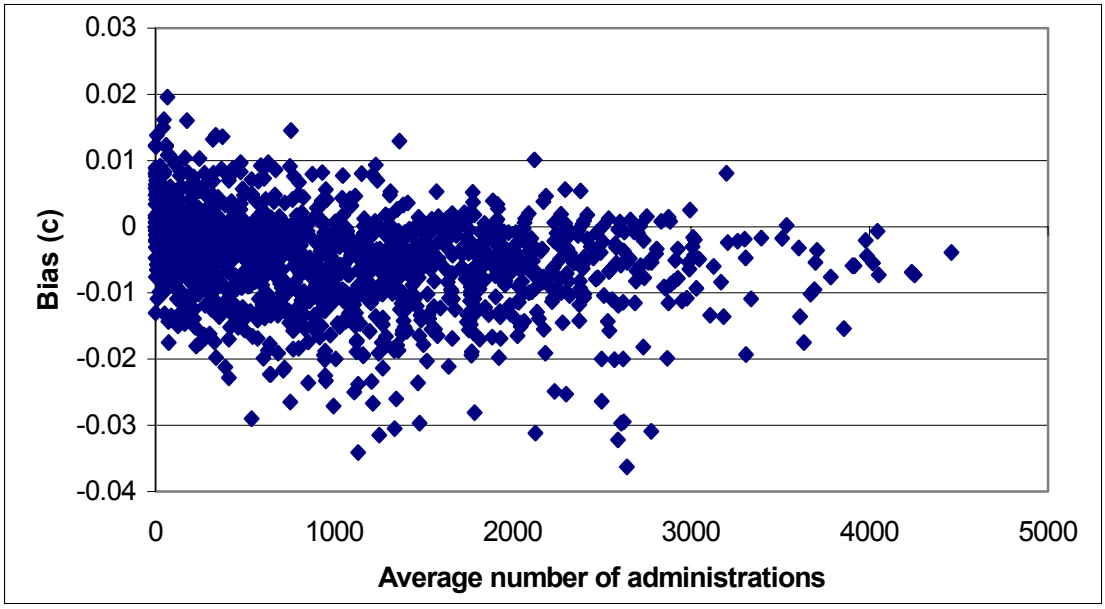


Figure 26. Scatterplot showing the relationship between the bias (generated –calibrated) for the c-parameter and the average number of administrations for an item.

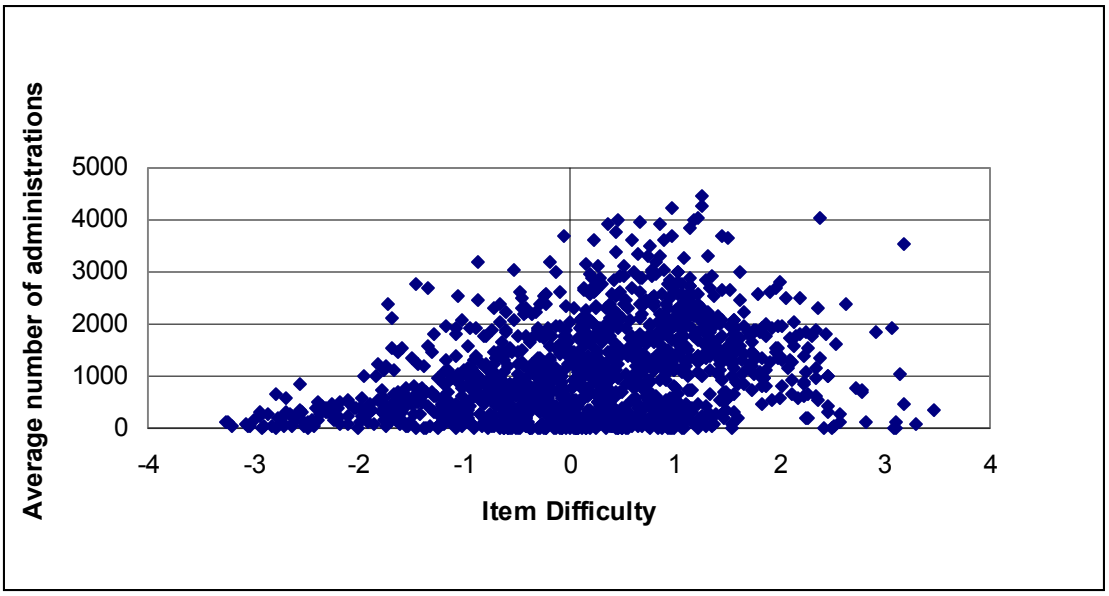


Figure 27. Scatterplot showing the relationship between the average number of administrations for an item (generated – calibrated) and item difficulty.