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

Iosif A. Krass
Personnel Testing Division, Defense Manpower Data Center
Seaside, California

Bruce Williams
ADAMS Inc, Urbana, Illinois

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

The views expressed are those of the authors and not necessarily those of the Department of Defense, or the United States Government.

In this paper we will present a study in which we used simulated data sets to consider and compare nonparametric and adjusted marginal maximum likelihood methods used to calibrate pretest items, as well as recalibrate items in the operational CAT pool. CAT data matrices are sparse because an examinee responds to a relatively small number of items over a narrow range of ability. Logistic parameters for conventional tests are commonly fit by direct marginal maximum likelihood fitting (MMLE) of logistic item response functions. It may be preferable to first fit a more general family of functions to the available data (by MMLE) and then indirectly obtain logistic parameters by fitting logistic curves to directly estimated more general functions. Some researchers have conjectured that even in logistic simulations, the additional degrees of freedom of a more general set of functions would make the initial model fitting program less sensitive to errors that inevitably occur when logistic parameters estimated over a narrow ability range are used to extrapolate to abilities far from the range.

Because CAT tests generate rather sparse test data sets, traditional calibration methods, which can be called direct methods, – Bilog-MG (Zimowski, Muraki, Mislevy, and Bock, 2003) and SPLITEM (Ramsay, 1996) – need some modification to get convergence to reasonable values. Together with traditional calibration tools which use a parametric item model – parametric Item Characteristic Curve (ICC) – we will consider a more general approach when using a calibration algorithm to initially search for the item model (item ICC) from a rather broad class of allowable models (nonparametric models). Then, if necessary, the other part of the calibration algorithm fits the parametric model to the estimated nonparametric model (indirect methods). The nonparametrc-indirect method (described below) uses Multilinear Formula Score Theory (MFS, Levine, 2003, in press) and a suite of model fitting programs collectively called ForScore (FS).

Brief Description of Considered Algorithms

Because the FS method is relatively new in calibration literature and is the centerpiece of this paper , we will focus the discussion on it; for the other methods that are already known in the psychometric world, we will describe only our modification. The direct approaches to logistic parameterization we considered are based on Bilog-MG Version 3.0 (Bilog-MG by M.Zimowsky, E. Muraki, R. Mislevy, D. Bock, 2002) and an MMLE logistic algorithm provided by J.O. Ramsay (1996). Both programs are designed for Paper and Pencil tests complete data matrices. We modified both programs to work-around adaptive data convergence problems. These two direct MMLE methods serve as benchmarks for estimating the performance of FS in this study.

We will consider unidimensional cases only, where every one of considered n examinees/simulees is completely characterized by the value of his/her latent ability q_i , $i = 1,\ldots,n$. Although FS can be easily transformed for multidimensional cases, in this paper we will stay with unidimensional cases. We assume that every considered item $j = 1,\ldots,m$ can be completely characterized by its ICC function $P_j(q)$, $j = 1,\ldots,m$; the value of that is the probability for an examinee with ability q to answer the item j right. Here ability belongs to the finite interval $q \in (q_{\min}, q_{\max})$, (e.g., in the case of the CAT-ASVAB tests, $q_{\min} = -3.0$ and $q_{\max} = 3.0$). We consider only binary answers (right, wrong) by an examinee for any item, although the FS method can easily handle more than one answer per item (graded response). Thus, in the considered case, the answer by an examinee to the item j can be characterized by the binary variable $u_j \in \{0,1\}$, where 0 corresponds to a wrong answer to item j, and 1 corresponds to the right answer.

If an examinee gets a test that consists of items $\{j_1, \ldots, j_k\}$, where $k \leq m$, then we will assume that answers on any two items in the test are independent events. If, as a result of the test, an examinee with ability q will "generate" a vector of answers $u = (u_{j_1}, \ldots, u_{j_k})$, the likelihood of the appearance of this vector is equal:

$$l(q,u) = \prod_{l=1}^{k} \{u_{j_l} \cdot P_{j_l}(q) + (1 - u_{j_l}) \cdot (1 - P_{j_l}(q))\}$$

due to assumed independence. Thus, for every binary vector u equal to, or less than, m, we get a smooth, real-value likelihood function. The space K spanned by those functions is called the canonical space, derived by items $j = 1, \ldots, m$. Because the number of the described-above binary vectors u is finite (e.g., if every considered test has a length m, then this number is 2^m), the canonical space is finite dimensional. We can introduce scalar product in canonical space as usual:

$$< l_1(q,u_1), l_2(q,u_2) >= \int_{q_{\min}}^{q_{\max}} l_1(q,u_1) \cdot l_2(q,u_2) \cdot dn \quad ,$$

where n is the density of latent ability distribution on interval $(q_{\min}, q_{\max})$. As it is shown in Caroll, Williams, and Levine (1999), this set can be extended to the closed subset in the Hilbert space $L^2(q_{\min}, q_{\max})$ of all integral functions on $(q_{\min}, q_{\max})$. Because the canonical space K is a finite dimensional linear space, it has an orthonormal basis $\{e_l(q)\}, l = 1, \ldots, J$, where J is the dimension of the canonical space, such that $< e_i(q), e_j(q) >= 0$, if $i \neq j$, and $< e_i(q), e_j(q) >= 1$, if $i = j$. Any function $p(q) \in \mathrm{K}$, can be presented in the form $p(q) = \sum_l c_l \cdot e_l(q)$, where $q \in (q_{\min}, q_{\max})$. Levine and Williams (1978) show that the orthonormal basis in K can be chosen in such a way that values $| c_l |$ of the above decomposition decrease rather rapidly with the growth of index l. Therefore, reasonable precision can be reached in an orthonormal presentation of arbitrary function with few local optimums in the interval $(q_{\min}, q_{\max})$, for $l \leq 14$ in the case of this special orthonormal basis. Thus, the problem of finding a new function which fits to the given test data set can be presented as a maximization-correspondent likelihood in a finite dimensional space, and it can be successfully solved (Caroll et al). The above maximization will produce a generally smooth, but nonparametric, item ICC.

In most applications, the item ICC should be presented in parametric form; for example, in the case of CAT-ASVAB, it should be presented as a 3PL item:

$$p_j(q) = c_j + \frac{1 - c_j}{1 + \exp(l_j(q))} \qquad ,$$

where $l_j(q) = -D \cdot a_j \cdot (q - b_j)$, and a_j, b_j, c_j are the item discriminating, difficulty, and guessing indices; $D = 1.7$ is a scaling constant (Lord, 1980); and examinee ability $q \in [q_{\min}, q_{\max}]$.

To fit a proper 3PL ICC to the chosen FS nonparametric curve $p(q) \in \mathrm{K}$, the finalizing part of FS uses a special part of a signal theory (Green and Swets, 1996) which connects with the

Independent Observer Index (IOI). From the mathematical point of view, this approach is very close to minimizing the quadratic function loss.

To compare the performance of some different approaches to the calibration problem, we chose two traditional methods – Bilog-MG, Version 3.0, and SPLITEM. Because both of these methods were designed for use with conventional (paper-and-pencil) tests (i. e., for nearly totally filled test matrix data sets), they have convergence problems and implausible fit problems with sparse CAT data sets. To get around this we used different approaches for these methods.

In the case of Bilog-MG, we are using a better initial state for maximum likelihood optimization than the originally designed prime state. To get this initial state, we solve the preliminary simulated paper-and-pencil problem for the operational CAT item pool by using the CAT estimation ability for the given set of examinees (so called theta-hat estimations). In recalibrating the operational CAT item pool, we are using the best available estimations of those parameters (prime estimations). (In calibrating the pretest items, we are using $(a = 1.0, b = 0.0, c = 0{,}5)$ as prime parameters because for those items, as usual, there is no available IRT information.) The solution to this paper-and-pencil problem is to use Bilog-MG in the initial phase (mechanism of IFNAME in Global statement of Bilog-MG script). This approach gives relatively good results, but sometimes it does not lead to convergence for all the calibrated pretest items (especially with the CAT-ASVAB technical tests). Another major difference between the discussed application of Bilog-MG and the use of this instrument as reported by M. Pommerich in this session (Pommerich & Segall, 2003) relates to using Newton-

Rafson after the proper number of EM iterations: M. Pommerich purposefully omitted this part of the Bilog-MG algorithm.

It is quite possible to use another approach to force Bilog-MG to be more stable (e.g., to use the construction described below in the RWlog algorithm), but we are staying with the method that was originally used by DMDC for the calibration of the pretest items that were being considered for the operational CAT pool. Actual calibration of pretest items should technically be done after recalibrating the CAT item pool, because, in the case of the pretest items, we are calibrating one item per run using the best estimations for the CAT item pool.

In the case of the SPLITEM method, we need to use the approach worked out by Levine and Williams (1998) to achieve the needed convergence and stability. They conjectured that the difficulty they encountered may have been due to the fact that there are few items to which both high ability and low ability examinees respond. In a series of experimental studies they observed that including a small number of simulated examinees responding to all of the items in the pool greatly increased the numerical stability and accuracy of maximum likelihood estimators, including FS. Stability and accuracy were substantially improved by adding as few as one simulated conventional examinee per hundred adaptive examinees. When calibrating operational items that had been previously seeded into an operational administration, RWlog (name of SPLITEM method with simulated addition) includes one-percent simulated conventional data along with the data for the adaptively administered items. The estimated logistic parameters used to select and score the items for the adaptive test are used to simulate the conventional item responses.

Numerical Results and Conclusion

To estimate the performance of the different methods, we are using a simulation CAT test data file that was done for unidimensional cases of 122,400 simulees with known ability distributions. We will describe in more detail the case of normally distributed simulees, because the results for the other cases are more or less close to those results. The item bank of the simulation CAT is split into four disjointed subsets, which we call CAT1, CAT2, CAT3, and CAT4. There are also 100 pretest items. A simulee gets a particular subset, for example CAT1, and a pretest item chosen out of the 100 pretest items. The simulee's CAT test consists of 15 items which were selected by CAT-ASVAB's selection mechanism (Segal, Moreno, Bloxom, and Hetter, 1997) that uses item information estimated at the point of the current ability of the simulee and is subjected to preliminary developed exposure-control parameters for the given CAT. We apply the usual Owen-Bayes update algorithm to estimate this CAT test ability for a simulee. As a result of this estimation, we also get the so-called theta-hat ($\hat{q}$) or Baysian estimation of the simulee's true ability.

In the total CAT item pool the different subsets (CAT1,…, CAT4) have 94, 137, 137, and 137 items, correspondingly. The test data file is done in such a way that 40,000 simulees get tests from CAT1, CAT2, or CAT3, and only 2,400 simulees get a test generated by the CAT4 item pool. The purpose of the CAT4 item pool is to estimate the means and standard deviations of the ability distribution of the given set of a simulee population to determine the need for any possible change to the test scale. Therefore, in our preliminary estimation of performance of the above methods, we will recalibrate only the CAT1, CAT2, and CAT3 item pools; thus, we have

120,000 simulees that “use” correspondent item pools (40,000 per separate CAT subset) and a total of 368 items in the entire CAT pool.

We are analyzing the performance of the different methods in two ways. First of all we are estimating how well they will estimate (“recover”) the value of the 3PL parameters of the estimated curves. This is especially important in the case of pretest items because the value of parameter a roughly defines the item information, and so determines if that item will be available to be selected for an operational CAT test; the value of parameter b provides a preliminary estimation of item difficulty.

On the other hand, it is well known that separate item parameter changes can compensate each other. Due to this, we estimate the closeness of the ICC curves using the Root Mean Square Error (RMSE), which we use very similarly to what was used by F. Drasgow (Hulin, Drasgow, and Parsons, 1983). To take into account the distribution of ability, we compute weights for the sum of RMSE based on the theta-hat distributions of the simulee population.

$$RMSE = \sqrt{\sum_i (P_1(q_i) - P_2(q_i))^2 \cdot w_i} \,,$$

where $q_i = q_{\min} + i \cdot \Delta;\ i = 0,\ldots,51; q_{\min} = -2.5,\ \Delta = 0.1$. Here $P_j(q_i),\ j = 1,2$ is the value of the correspondent ICC at point q_i of the chosen grid, and w_i is the weight assigned to this point. To get those weights we build a histogram of theta-hats of the population using the above grid, and normalize the values of the histogram such that $\sum_i w_i = N$, where N is the number of points in our grid. In this case, we can compare our RMSE estimation with RMSE values reported by other researchers. The described method of RMSE-weighting allows us to emphasize the comparison of ICCs in the area of the ability interval that is more populated.

This estimation was used in the initial calibration of pretest items where theta-hat was the only available estimation of examinee ability. Note, however, that it is possible to use the RSME method, proposed by D. Segall and discussed here by him and M. Pommerich (Pommerich & Segall, 2003), based on knowledge of true ability of the correspondent simulee. With this approach, they take into account only those simulees with known true ability who got this item in their test, not the entire population of simulees. For those simulees with known true ability, the absolute differences between the two estimated ICCs are computed and provide part in RMSE summation. For that reason, numerical values of RMSE presented in this paper are, in general, higher than the values of RMSE presented by Pommerich and Segall. Note that A. Nicewander uses their method of RMSE for FS in the final paper in today's general conclusion (Nicewander, 2003).

We will stop with more detail in the case of a normal $N(0,1)$ distribution of a simulee's ability, because the analysis of the other two cases $N(1,\,0.8)$ and $N(-1,1.2)$ looks analogous. It is worth while to remark that in the case of the "shifted" distributions, $N(1,\,0.8)$, and $N(-1,1.2)$, the set of recovered parameters produced Bilog-MG, as well as by RWlog, essentially benefited from a Stocking-Lord transformation (Stocking & Lord, 1983) which is used to put ability estimates on a common scale. If $(a_0\,, b_{0\,,} c_0)$ are parameter values after calibration, then the values of the Stocking-Lord transformed parameters $(a, b_, c)$ are

$b = E(\hat{q}) + ST(\hat{q}) \cdot b_0$; $a = {a_0}/{ST(\hat{q})}$; $c = c_0$, where $E(\hat{q}), ST(\hat{q})$ are the means and standard deviations of ability distributions of the correspondent simulee population.

FS does not require the application of this transformation because the FS process begins from the estimation of the population ability distribution. More importantly, one of the FS outputs shows the ability distribution for the given examinee population.

Results for Calibration of Pretest Items (Figures 1 and 2)

In Figure 1 we show the frequencies of absolute values for all three methods. Even though we present results of other algorithms, remember that our major focus is on FS. These diagrams can be used to make a rough estimation of the precision of each method in two ways. For example, from the diagram for FS in Figure 1, it follows that for 40% of all the calibrated items, the absolute difference between the true and estimated parameter b is not more than 0.05. On the other hand, the frequency (chance) to get absolute values for difficulty parameters more than 0.3 is less than 0.1. In other words, the lower the correspondent curve for the upper part of the values of absolute deviation, and the closer to the y axes for the lower part of the values of the absolute deviation values, the better the correspondent estimator.

As we can see, the deviations of Bilog-MG look as good as the deviations for FS, but Bilog-MG does not converge for 56 pretest items. In the other words, only when Bilog-MG converges, does it produce rather good estimations.

Table 1 provides the means and standard deviations for the estimated parameters in the case of the pretest items.

Table 1. Pretest Items: means and standard deviations for estimated parameters.

	MEAN			STD		
Products	delta a	delta b	delta c	delta a	delta b	delta c
FS	0.253	0.108	0.045	0.22	0.12	0.05
RWlog	0.522	0.192	0.052	0.43	0.178	0.052
BilogMG	0.236	0.108	0.028	0.19	0.101	0.026

In Figure 2 we show graphs of the weighted RMSE deviations for the estimated ICCs for the different methods. Again, Bilog-MG looks as good as FS if we do not consider the convergence problems which FS or RWlog do not have.

Results for Recalibration of Item Pool (Figures 3 and 4)

All the figures and tables below use data aggregated from the CAT1, CAT2, and CAT3 item pools. Figure 3 shows analogous graphs for the recalibration of the items in the CAT pool. Once again, Bilog-MG was able to make estimations for all except 1 of the 368 calibrated items, though its estimations are not quite as good as the other two products. Table 2 shows the means and standard deviations for the estimated parameters in the case of the CAT pool recalibration.

Table 2. Item Pool Recalibration: means and standard deviations for estimated parameters.

	MEAN			STD		
Products	delta a	delta b	delta c	delta a	delta b	delta c
FS	0.104	0.056	0.03	0.103	0.074	0.042
RWlog	0.126	0.072	0.029	0.105	0.089	0.035
Bilog-MG	0.231	0.111	0.043	0.217	0.142	0.035

Figure 4 presents graph frequencies for RMSE deviations in the CAT item pool recalibration. As we can see, FS looks considerably better than the other two products, and RWlog is slightly better than Bilog-MG.. Table 3 provides the means and standard deviations for RMSE for both cases: pretest items and items in the CAT pool.

Table 3. Pretest Items and Item Pool: means and standard deviations for RMSE for three different methods.

Type	Pretest Items		Item Pool	
Statistic	Mean	STD	Mean	STD
FS	0.065	0.033	0.024	0.012
RWlog	0.156	0.086	0.035	0.024
BilogMG	0.064	0.032	0.085	0.073

As we can see again, Bilog-MG is has the same precision as FS in the case of the pretest item recovery, but it can estimate only 56 pretest items out of 100.

Overall, in the case of normal-normal distributed simulees $N(0, 1)$, FS appears more precise and stable. The “second” place in this “competition” belongs, in our opinion, to the RWlog algorithm, which is considerably more stable than Bilog-MG.

With simulees that have a shifted distribution to the left $N(-1, 1.2)$, (case of “less able” simulees) we get about the same results with FS, RWlog, and Bilog-MG. In this case, Bilog-MG, using the same scheme of choosing the better initial state for running, cannot recalibrate the items in the CAT pool (computer computation blew up in attempt of take logarithm of negative number in time of internal iteration). The same thing happened when the population of simulees was shifted to the right: case of $N(1, 0.8)$ distribution.

In the case of a “less able”simulee population$N(-1, 1.2)$, Bilog-MG cannot estimate 62 out of 100 pretest items. In the case of a “more able” simulee population , $N(1, 0.8)$ RWlog makes a more precise estimation (recovery), especially parameter-wise than FS as in the pretest item case, as well as in item pool recovery case. As we already mentioned, Bilog-MG was not able to recalibrate the CAT pool and could not calibrate 14 seeded items out of 100. Thus, the winner of this "competition" in the case of $N(1, 0.8)$ is RWlog. In all other case the winner is FS. This phenomena was amazing, because in very many preliminary simulated and not-simulated runs (at least in 100 cases), FS was always much more precise than RWlog. Considering this in more detail, we found that both less-able and more-able cases $N(-1, 1.2)$ and $N(1, 0.8)$ are far from reality. In all real examples and our preliminary simulations, when we try to imitate real cases, the maximum shift of mean of ability distribution was not more than 0.5 by absolute value. The drastic changes of ability as $N(1, 0.8)$ or $N(-1, 1.2)$ require a special tune up of the FS part that is responsible for the estimation of ability distribution, which can be done, if necessary.

References

Caroll J. D., Williams B., & Levine M. V. (in press). Recovering a multidimensional model from fitted unidimensional submodels. *British Journal of Mathematical and Statistical Psychology*.

Green, D., & Swets, J. (1996). *Signal detection theory and psychophysics.* New York: Willey.

Hulin, C. L., Drasgow, F., & Parsons, C. K. (1983). *Item response theory. Application to psychological measurement*. Homewood, IL: Dow Jones-Irvin.

Levine M. V. (in press). Dimension in latent variable models. *Journal of Mathematical Psychology*.

Lord, F.M. (1980). *Applications of IRT to practical problems.* Hillsdale: LEA.

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.

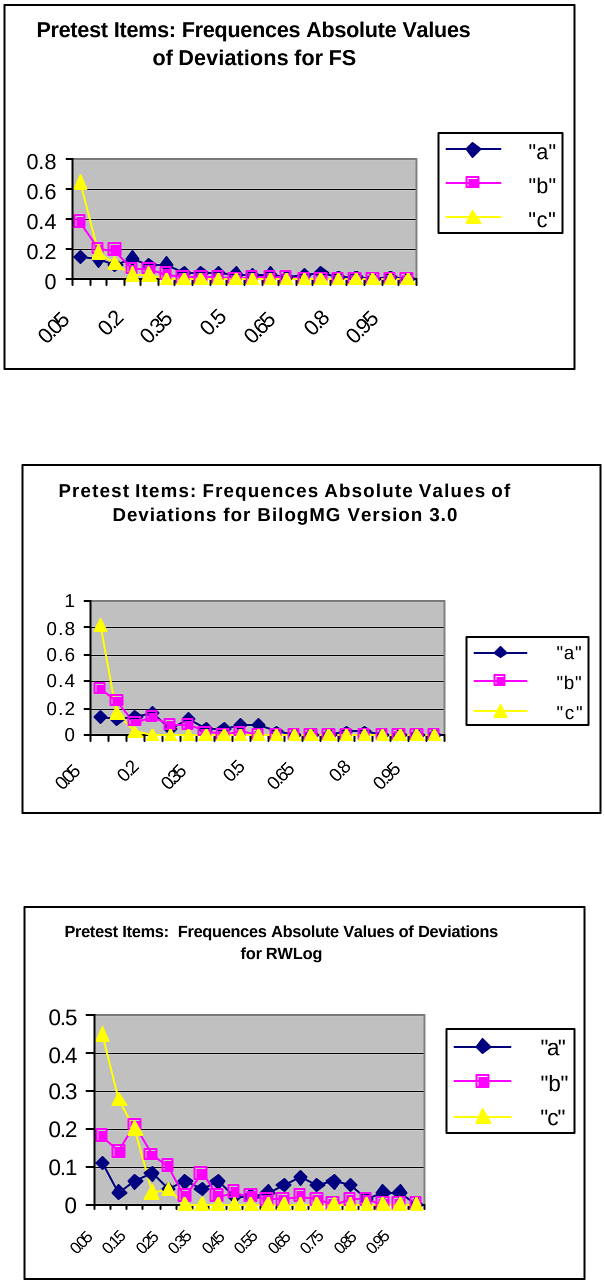
Pommerich, M., & Segall, D. O. (2003). *Calibrating CAT Pools and Online Pretest Items Using Marginal Maximum Likelihood Methods*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Chicago, IL.

Ramsay, J. O. (1996). A functional approach to modeling test data. Wim J. van der Linden *Handbook of Modern Item Response Theory*. New York, London: Springer.

Segall, D. O., Moreno, K. E., Bloxom, B. M., & Hetter, R. D. (1997). Psychometric procedures for administering CAT-ASVAB. In W. A. Sands, B. K. Waters, & J. R. McBride (Eds.), *Computerized Adaptive Testing* (pp. 131-140). Washington, D C: American Psychological Association.

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

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.


Pretest Items: Frequences Absolute Values of Deviations for FS

Pretest Items: Frequences Absolute Values of Deviations for BilogMG Version 3.0

Pretest Items: Frequences Absolute Values of Deviations for RWLog

Figure 1. Pretest items: Frequencies of deviations of absolute values for three different calibration methods.

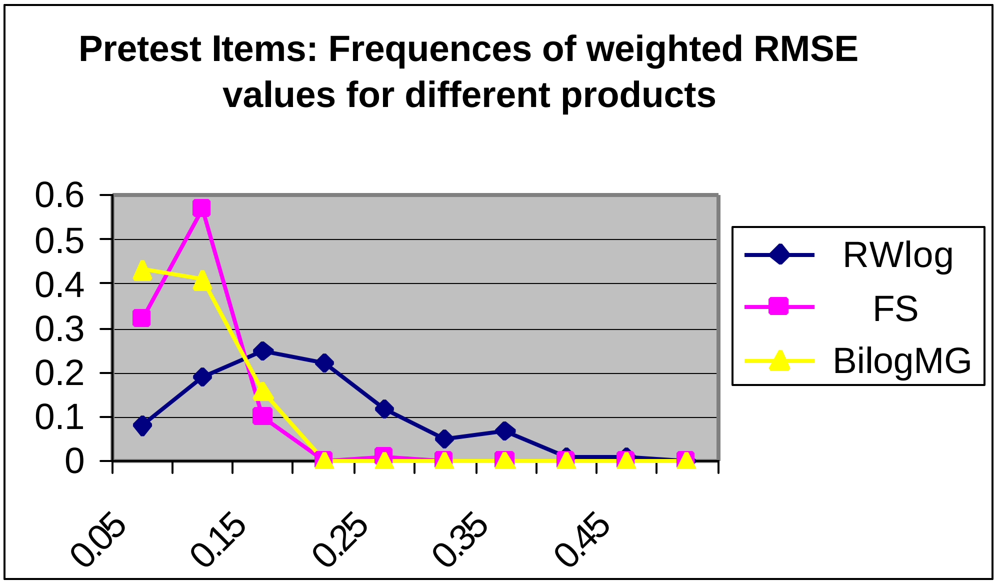
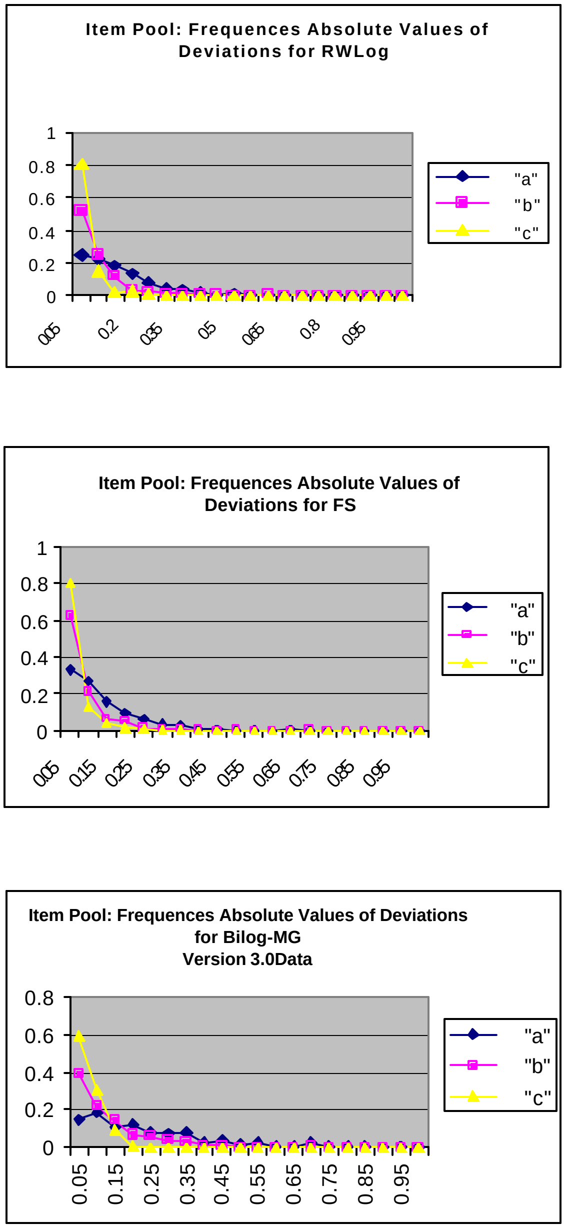


Figure 2. Pretest Items: Frequencies of RMSE values for three different methods.



Item Pool: Frequences Absolute Values of Deviations for RWLog

Item Pool: Frequences Absolute Values of Deviations for FS

Item Pool: Frequences Absolute Values of Deviations for Bilog-MG Version 3.0Data

Figure 3. Item Pool Recalibration: Frequencies of deviations of absolute values for three different calibration methods.

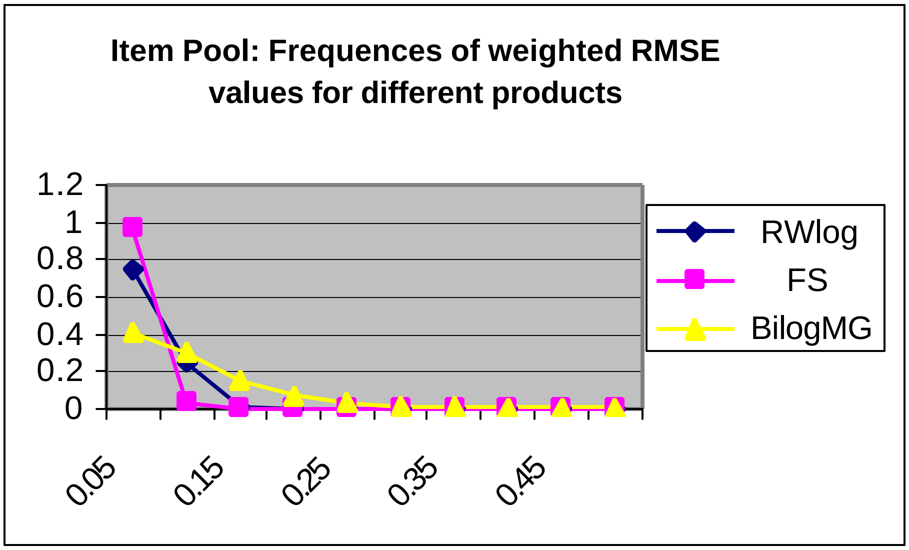


Figure 4. Item Pool Recalibration: Frequencies of RMSE values for three different methods.

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

Iosif A. Krass
Personnel Testing Division, Defense Manpower Data Center
Seaside, California

Bruce Williams
ADAMS Inc, Urbana, Illinois

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

The views expressed are those of the authors and not necessarily those of the Department of Defense, or the United States Government.

In this paper we will present a study in which we used simulated data sets to consider and compare nonparametric and adjusted marginal maximum likelihood methods used to calibrate pretest items, as well as recalibrate items in the operational CAT pool. CAT data matrices are sparse because an examinee responds to a relatively small number of items over a narrow range of ability. Logistic parameters for conventional tests are commonly fit by direct marginal maximum likelihood fitting (MMLE) of logistic item response functions. It may be preferable to first fit a more general family of functions to the available data (by MMLE) and then indirectly obtain logistic parameters by fitting logistic curves to directly estimated more general functions. Some researchers have conjectured that even in logistic simulations, the additional degrees of freedom of a more general set of functions would make the initial model fitting program less sensitive to errors that inevitably occur when logistic parameters estimated over a narrow ability range are used to extrapolate to abilities far from the range.

Because CAT tests generate rather sparse test data sets, traditional calibration methods, which can be called direct methods, – Bilog-MG (Zimowski, Muraki, Mislevy, and Bock, 2003) and SPLITEM (Ramsay, 1996) – need some modification to get convergence to reasonable values. Together with traditional calibration tools which use a parametric item model – parametric Item Characteristic Curve (ICC) – we will consider a more general approach when using a calibration algorithm to initially search for the item model (item ICC) from a rather broad class of allowable models (nonparametric models). Then, if necessary, the other part of the calibration algorithm fits the parametric model to the estimated nonparametric model (indirect methods). The nonparametrc-indirect method (described below) uses Multilinear Formula Score Theory (MFS, Levine, 2003, in press) and a suite of model fitting programs collectively called ForScore (FS).

Brief Description of Considered Algorithms

Because the FS method is relatively new in calibration literature and is a center piece of this paper , we will focus the discussion on it; for the other methods that are already known in the psychometric world, we will describe only our modification. The direct approaches to logistic parameterization we considered are based on Bilog-MG Version 3.0 (Bilog-MG by M.Zimowsky, E. Muraki, R. Mislevy, D. Bock, 2002) and an MMLE logistic algorithm provided by J.O. Ramsay (1996). Both programs are designed for Paper and Pencil tests complete data matrices. We modified both programs to work-around adaptive data convergence problems. This two direct MMLE methods serve for this research as a benchmark for estimation performance of FS.

We will consider unidimensional cases only, where every one of considered n examinees/simulees is completely characterized by the value of his/her latent ability q_i , $i = 1, \ldots, n$. Although FS can be easily transformed for multidimensional cases, in this paper we will stay with unidimensional cases. We assume that every considered item $j = 1, \ldots, m$ can be completely characterized by its ICC function $P_j(q)$, $j = 1, \ldots, m$; the value of that is the probability for an examinee with ability q to answer the item j right. Here ability belongs to the finite interval $q \in (q_{\min}, q_{\max})$, (e.g., in the case of the CAT-ASVAB tests, $q_{\min} = -3.0$ and $q_{\max} = 3.0$). We consider only binary answers (right, wrong) by an examinee for any item, although the FS method can easily handle more than one answer per item (graded response). Thus, in the considered case, the answer by an examinee to the item j can be characterized by the binary variable $u_j \in \{0,1\}$, where 0 corresponds to a wrong answer to item j, and 1 corresponds to the right answer.

If an examinee gets a test that consists of items $\{j_1, \ldots, j_k\}$, where $k \le m$, then we will assume that answers on any two items in the test are independent events. If, as a result of the test, an examinee with ability q will "generate" a vector of answers $u = (u_{j_1}, \ldots, u_{j_k})$, the likelihood of the appearance of this vector is equal:

$$l(q,u) = \prod_{l=1}^{k} \{u_{j_l} \cdot P_{j_l}(q) + (1 - u_{j_l}) \cdot (1 - P_{j_l}(q))\}$$

due to assumed independence. Thus, for every binary vector u equal to, or less than, m, we get a smooth, real-value likelihood function. The space K spanned by those functions is called the canonical space, derived by items $j = 1, \ldots, m$. Because the number of the described-above binary vectors u is finite (e.g., if every considered test has a length m, then this number is 2^m), the canonical space is finite dimensional. We can introduce scalar product in canonical space as usual:

$$< l_1(q,u_1), l_2(q,u_2) >= \int_{q_{\min}}^{q_{\max}} l_1(q,u_1) \cdot l_2(q,u_2) \cdot dn \quad ,$$

where n is the density of latent ability distribution on interval $(q_{\min}, q_{\max})$. As it is shown in Caroll, Williams, and Levine (1999), this set can be extended to the closed subset in the Hilbert space $L^2(q_{\min}, q_{\max})$ of all integral functions on $(q_{\min}, q_{\max})$. Because the canonical space K is a finite dimensional linear space, it has an orthonormal basis $\{e_l(q)), l = 1, \ldots, J$, where J is the dimension of the canonical space, such that $< e_i(q), e_j(q) >= 0$, if $i \neq j$, and $< e_i(q), e_j(q) >= 1$, if $i = j$. Any function $p(q) \in \mathrm{K}$, can be presented in the form $p(q) = \sum_l c_l \cdot e_l(q)$, where $q \in (q_{\min}, q_{\max})$. Levine and Williams (1978) show that the orthonormal basis in K can be chosen in such a way that values $| c_l |$ of the above decomposition decrease rather rapidly with the growth of index l. Therefore, reasonable precision can be reached in an orthonormal presentation of arbitrary function with few local optimums in the interval $(q_{\min}, q_{\max})$, for $l \leq 14$ in the case of this special orthonormal basis. Thus, the problem of finding a new function which fits to the given test data set can be presented as a maximization-correspondent likelihood in a finite dimensional space, and it can be successfully solved (Caroll et al). The above maximization will produce a generally smooth, but nonparametric, item ICC.

In most applications, the item ICC should be presented in parametric form; for example, in the case of CAT-ASVAB, it should be presented as a 3PL item:

$$p_j(q) = c_j + \frac{1 - c_j}{1 + \exp(l_j(q))} \qquad ,$$

where $l_j(q) = -D \cdot a_j \cdot (q - b_j)$, and a_j, b_j, c_j are the item discriminating, difficulty, and guessing indices; $D = 1.7$ is a scaling constant (Lord, 1980); and examinee ability $q \in [q_{\min}, q_{\max}]$.

To fit a proper 3PL ICC to the chosen FS nonparametric curve $p(q) \in \mathrm{K}$, the finalizing part of FS uses a special part of a signal theory (Green and Swets, 1996) which connects with the

Independent Observer Index (IOI). From the mathematical point of view, this approach is very close to minimizing the quadratic function loss.

To compare the performance of some different approaches to the calibration problem, we chose two traditional methods – Bilog-MG, Version 3.0, and SPLITEM. Because both of these methods were designed for use with conventional (paper-and-pencil) tests (i. e., for nearly totally filled test matrix data sets), they have convergence problems and implausible fit problems with sparse CAT data sets. To get around this we used different approaches for these methods.

In the case of Bilog-MG, we are using a better initial state for maximum likelihood optimization than the originally designed prime state. To get this initial state, we solve the preliminary simulated paper-and-pencil problem for the operational CAT item pool by using the CAT estimation ability for the given set of examinees (so called theta-hat estimations). In recalibrating the operational CAT item pool, we are using the best available estimations of those parameters (prime estimations). (In calibrating the pretest items, we are using $(a = 1.0, b = 0.0, c = 0{,}5)$ as prime parameters because for those items, as usual, there is no available IRT information.) The solution to this paper-and-pencil problem is to use Bilog-MG in the initial phase (mechanism of IFNAME in Global statement of Bilog-MG script). Another major difference between the discussed application of Bilog-MG and usage of this instrument as reported by Mary Pommerich in this session consists of using Newton- Rafson after proper number of EM iteration. Mary is totally eliminated this part of Bilog_mg algorithm. This approach gives relatively good results, but sometimes it does not lead to convergence for all the calibrated pretest items (especially with the CAT-ASVAB technical tests).

It is quite possible to use another approach to force Bilog-MG to be more stable (e.g., to use the construction described below in the RWlog algorithm), but we are staying with the method that was originally used by DMDC for the calibration of the pretest items that were being considered for the operational CAT pool. Actual calibration of pretest items should technically be done after recalibrating the CAT item pool, because, in the case of the pretest items, we are calibrating one item per run using the best estimations for the CAT item pool.

In the case of the SPLITEM method, we need to use the approach worked out by Levine and Williams (1998) to achieve the needed convergence and stability. They conjectured that the difficulty they encountered may have been due to the fact that there are few items to which both high ability and low ability examinees respond. In a series of experimental studies they observed that including a small number of simulated examinees responding to all of the items in the pool greatly increased the numerical stability and accuracy of maximum likelihood estimators, including FS. Stability and accuracy were substantially improved by adding as few as one simulated conventional examinee per hundred adaptive examinees. When calibrating operational items that had been previously seeded into an operational administration, RWlog (name of SPLITEM method with simulated addition) includes one-percent simulated conventional data along with the data for the adaptively administered items. The estimated logistic parameters used to select and score the items for the adaptive test are used to simulate the conventional item responses.

Numerical Results and Conclusion

To estimate the performance of the different methods, we are using a simulation CAT test data file that was done for unidimensional cases of 122,400 simulees with known ability distributions. We will describe in more detail the case of normally distributed simulees, because the results for the other cases are more or less close to those results. The item bank of the simulation CAT is split into four disjointed subsets, which we call CAT1, CAT2, CAT3, and CAT4. There are also 100 pretest items. A simulee gets a particular subset, for example CAT1, and a pretest item chosen out of the 100 pretest items. The simulee's CAT test consists of 15 items which were selected by CAT-ASVAB's selection mechanism (Segal, Moreno, Bloxom, and Hetter, 1997) that uses item information estimated at the point of the current ability of the simulee and is subjected to preliminary developed exposure-control parameters for the given CAT. We apply the usual Owen-Bayes update algorithm to estimate this CAT test ability for a simulee. As a result of this estimation, we also get the so-called theta-hat ($\hat{q}$) or Baysian estimation of the simulee's true ability.

In the total CAT item pool the different subsets (CAT1,…, CAT4) have 94, 137, 137, and 137 items, correspondingly. The test data file is done in such a way that 40,000 simulees get tests from CAT1, CAT2, or CAT3, and only 2,400 simulees get a test generated by the CAT4 item pool. The purpose of the CAT4 item pool is to estimate the means and standard deviations of the ability distribution of the given set of a simulee population to determine the need for any possible change to the test scale. Therefore, in our preliminary estimation of performance of the above methods, we will recalibrate only the CAT1, CAT2, and CAT3 item pools; thus, we have

120,000 simulees that “use” correspondent item pools (40,000 per separate CAT subset) and a total of 368 items in the entire CAT pool.

We are analyzing the performance of the different methods in two ways. First of all we are estimating how well they will estimate (“recover”) the value of the 3PL parameters of the estimated curves. This is especially important in the case of pretest items because the value of parameter a roughly defines the item information, and so determines if that item will be available to be selected for an operational CAT test; the value of parameter b provides a preliminary estimation of item difficulty.

On the other hand, it is well known that separate item parameter changes can compensate each other. Due to this, we estimate the closeness of the ICC curves using the Root Mean Square Error (RMSE), which we use very similarly to what was used by F. Drasgow (Hulin, Drasgow, and Parsons, 1983). To take into account the distribution of ability, we compute weights for the sum of RMSE based on the theta-hat distributions of the simulee population.

$$RMSE = \sqrt{\sum_i (P_1(q_i) - P_2(q_i))^2 \cdot w_i} \ ,$$

where $q_i = q_{\min} + i \cdot \Delta; i = 0, \ldots, 51; q_{\min} = -2.5, \Delta = 0.1$. Here $P_j(q_i), j = 1,2$ is the value of the correspondent ICC at point q_i of the chosen grid, and w_i is the weight assigned to this point. To get those weights we build a histogram of theta-hats of the population using the above grid, and normalize the values of the histogram such that $\sum_i w_i = N$, where N is the number of points in our grid. In this case, we can compare our RMSE estimation with RMSE values reported by other researchers. The described method of RMSE-weighting allows us to emphasize the comparison of ICCs in the area of the ability interval that is more populated.

This estimation was used in the initial calibration of pretest items where theta-hat is the only available estimation of examinee ability. However, in this presented simulation comparison it is possible to use the RSME method proposed by D. Segal and discussed here by him and M. Pommerich (Pommerich and Segall, 2003) based on knowledge of true ability of the correspondent simulee. In they approach for the given item they take in account not all population of examinees but only those examinee who got this item in their test. For those examinees with the known true ability absolute differences between two estimated ICC computed and providing its part in RMSE summation. For this reason numerical values of RMSE presented in they papers in general more hiher than the values of RMSE presentedby Segal and Pommerich. A. Nicewander using they method of RMSE for FS in the final paper in today's general conclusion (Nicewander, 2003).

We will stop with more detail in the case of a normal $N(0,1)$ distribution of a simulee's ability, because the analysis of the other two cases $N(1, 0.8)$ and $N(-1, 1.2)$ looks analogous. It is worth while to remark that in the case of the "shifted" distributions, $N(1, 0.8)$, and $N(-1, 1.2)$, the set of recovered parameters produced Bilog-MG, as well as by RWlog, essentially benefited from a Stocking-Lord transformation (Stocking & Lord, 1983) which is used to put ability estimates on a common scale. If $(a_0, b_{0,} c_0)$ are parameter values after calibration, then the values of the Stocking-Lord transformed parameters $(a, b_, c)$ are

$b = E(\hat{q}) + ST(\hat{q}) \cdot b_0$; $a = {}^{a_0}\!/_{ST(\hat{q})}$; $c = c_0$, where $E(\hat{q}), ST(\hat{q})$ are the means and standard deviations of ability distributions of the correspondent simulee population.

FS does not require the application of this transformation because the FS process begins from the estimation of the population ability distribution. More importantly, one of the FS outputs shows the ability distribution for the given examinee population.

Results for Calibration of Pretest Items (Figures 1 and 2)

Even we presented results of other algorithms here our major focus is FS. In Figure 1 we show the frequencies of absolute values for all three methods. These diagrams can be used to make a rough estimation of the precision of each method in two ways. For example, from the diagram for FS in Figure 1, it follows that for 40% of all the calibrated items, the absolute difference between the true and estimated parameter b is not more than 0.05. On the other hand, the frequency (chance) to get absolute values for difficulty parameters more than 0.3 is less than 0.1. In other words, the lower the correspondent curve for the upper part of the values of absolute deviation, and the closer to the y axes for the lower part of the values of the absolute deviation values, the better the correspondent estimator.

As we can see, the deviations of Bilog-MG look as good as the deviations for FS, but Bilog-MG does not converge for 56 pretest items. In the other words, only when Bilog-MG converges, does it produce rather good estimations.

Table 1 provides the means and standard deviations for the estimated parameters in the case of the pretest items.

Table 1. Pretest Items: means and standard deviations for estimated parameters.

	MEAN			STD		
Products	delta a	delta b	delta c	delta a	delta b	delta c
FS	0.253	0.108	0.045	0.22	0.12	0.05
RWlog	0.522	0.192	0.052	0.43	0.178	0.052
BilogMG	0.236	0.108	0.028	0.19	0.101	0.026

In Figure 2 we show graphs of the weighted RMSE deviations for the estimated ICCs for the different methods. Again, Bilog-MG looks as good as FS if we do not consider the convergence problems which FS or RWlog do not have.

Results for Recalibration of Item Pool (Figures 3 and 4)

All the figures and tables below use data aggregated from the CAT1, CAT2, and CAT3 item pools. Figure 3 shows analogous graphs for the recalibration of the items in the CAT pool. Once again, Bilog-MG was able to make estimations for all except 1 of the 368 calibrated items, though its estimations are not quite as good as the other two products. Table 2 shows the means and standard deviations for the estimated parameters in the case of the CAT pool recalibration.

Table 2. Item Pool Recalibration: means and standard deviations for estimated parameters.

	MEAN			STD		
Products	delta a	delta b	delta c	delta a	delta b	delta c
FS	0.104	0.056	0.03	0.103	0.074	0.042
RWlog	0.126	0.072	0.029	0.105	0.089	0.035
Bilog-MG	0.231	0.111	0.043	0.217	0.142	0.035

Figure 4 presents graph frequencies for RMSE deviations in the CAT item pool recalibration. As we can see, FS looks considerably better than the other two products, and RWlog is slightly better than Bilog-MG.. Table 3 provides the means and standard deviations for RMSE for both cases: pretest items and items in the CAT pool.

Table 3. Pretest Items and Item Pool: means and standard deviations for RMSE for three different methods.

Type	Pretest Items		Item Pool	
Statistic	Mean	STD	Mean	STD
FS	0.065	0.033	0.024	0.012
RWlog	0.156	0.086	0.035	0.024

BilogMG	0.064	0.032	0.085	0.073

As we can see again, Bilog-MG is has the same precision as FS in the case of the pretest item recovery, but it can estimate only 56 pretest items out of 100.

Overall, in the case of normal-normal distributed simulees $N(0, 1)$, FS appears more precise and stable. The “second” place in this “competition” belongs, in our opinion, to the RWlog algorithm, which is considerably more stable than Bilog-MG.

With simulees that have a shifted distribution to the left $N(-1, 1.2)$, (case of “less able” simulees) we get about the same results with FS, RWlog, and Bilog-MG. In this case, Bilog-MG, using the same scheme of choosing the better initial state for running, cannot recalibrate the items in the CAT pool (computer computation blew up in attempt of take logarithm of negative number in time of internal iteration). The same thing happened when the population of simulees was shifted to the right: case of $N(1, 0.8)$ distribution.

In the case of a “less able”simulee population$N(-1, 1.2)$, Bilog-MG cannot estimate 62 out of 100 pretest items. In the case of a “more able” simulee population , $N(1, 0.8)$ RWlog makes a more precise estimation (recovery), especially parameter-wise than FS as in the pretest item case, as well as in item pool recovery case. As we already mentioned, Bilog-MG was not able to recalibrate the CAT pool and could not calibrate 14 seeded items out of 100. Thus, the winner of this “competition” in the case of $N(1, 0.8)$ is RWlog. In all other case the winner is FS. This phenomena was amazing, because in very many preliminary simulated and not-simulated runs (at least in 100 cases), FS was always much more precise than RWlog. Considering this in more detail, we found that both less-able and more-able cases $N(-1, 1.2)$ and $N(1, 0.8)$ are far from reality. In all real examples and our preliminary simulations, when we try to imitate real cases, the maximum shift of mean of ability distribution was not more than 0.5 by absolute value. The drastic changes of ability as $N(1, 0.8)$ or $N(-1, 1.2)$ require a special tune up of the FS part that is responsible for the estimation of ability distribution, which can be done, if necessary.

References

Caroll J. D., Williams B., & Levine M. V. (in press). Recovering a multidimensional model from fitted unidimensional submodels. *British Journal of Mathematical and Statistical Psychology*.

Green, D., & Swets, J. (1996). *Signal detection theory and psychophysics.* New York: Willey.

Hulin, C. L., Drasgow, F., & Parsons, C. K. (1983). *Item response theory. Application to psychological measurement*. Homewood, IL: Dow Jones-Irvin.

Levine M. V. (in press). Dimension in latent variable models. *Journal of Mathematical Psychology.*

Lord, F.M. (1980). *Applications of IRT to practical problems.* Hillsdale: LEA.

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.

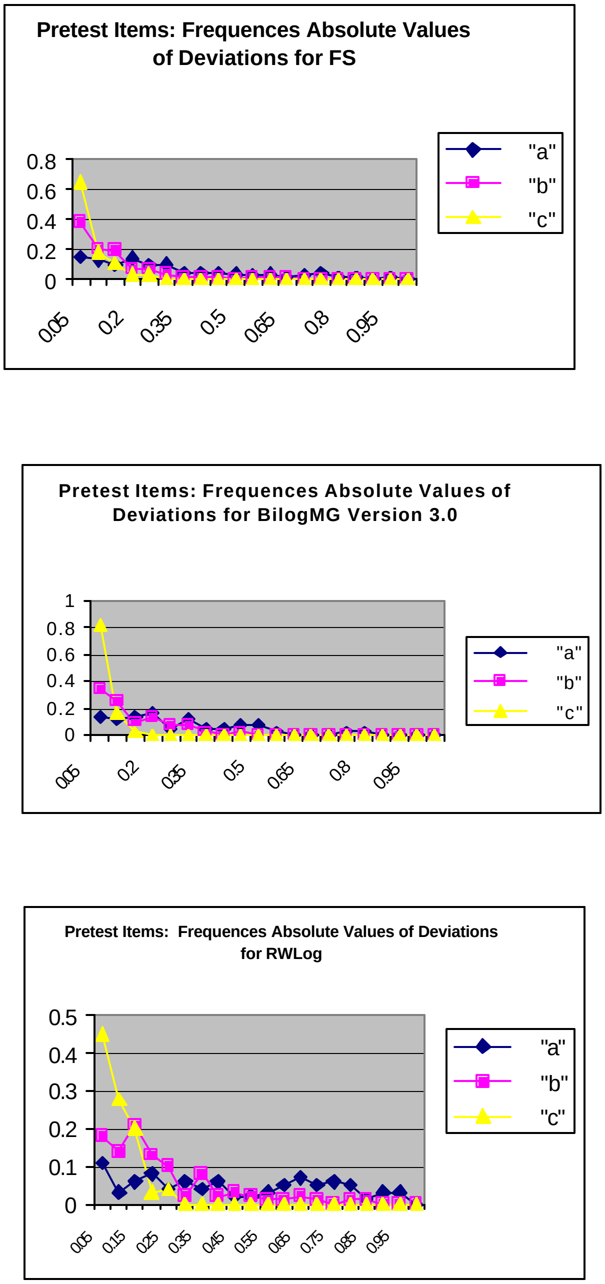
Pommerich, M., & Segall, D. O. (2003). *Calibrating CAT Pools and Online Pretest Items Using Marginal Maximum Likelihood Methods*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Chicago, IL.

Ramsay, J. O. (1996). A functional approach to modeling test data. Wim J. van der Linden *Handbook of Modern Item Response Theory*. New York, London: Springer.

Segall, D. O., Moreno, K. E., Bloxom, B. M., & Hetter, R. D. (1997). Psychometric procedures for administering CAT-ASVAB. In W. A. Sands, B. K. Waters, & J. R. McBride (Eds.), *Computerized Adaptive Testing* (pp. 131-140). Washington, D C: American Psychological Association.

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

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.



Pretest Items: Frequences Absolute Values of Deviations for FS

Pretest Items: Frequences Absolute Values of Deviations for BilogMG Version 3.0

Pretest Items: Frequences Absolute Values of Deviations for RWLog

Figure 1. Pretest items: Frequencies of deviations of absolute values for three different calibration methods.

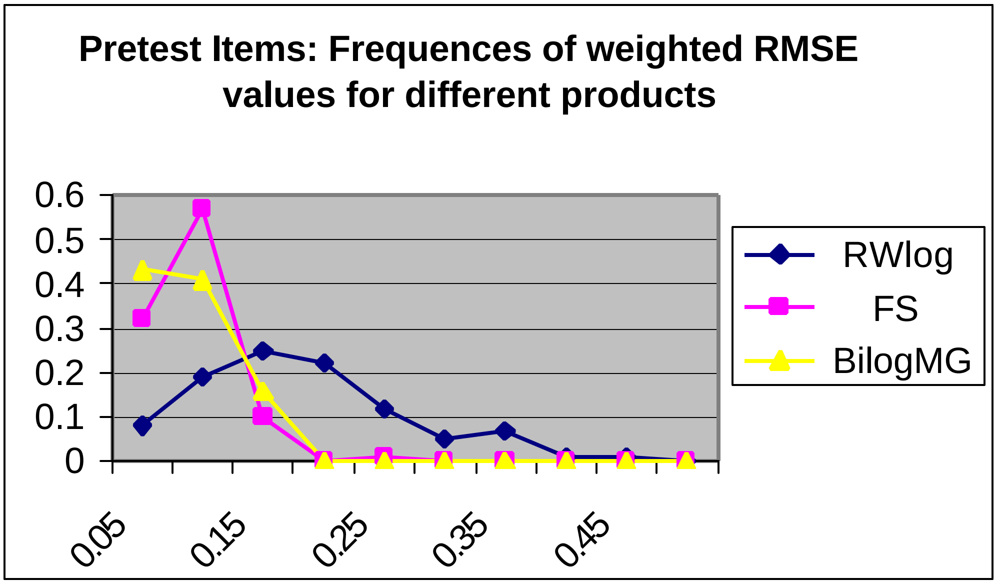


Figure 2. Pretest Items: Frequencies of RMSE values for three different methods.

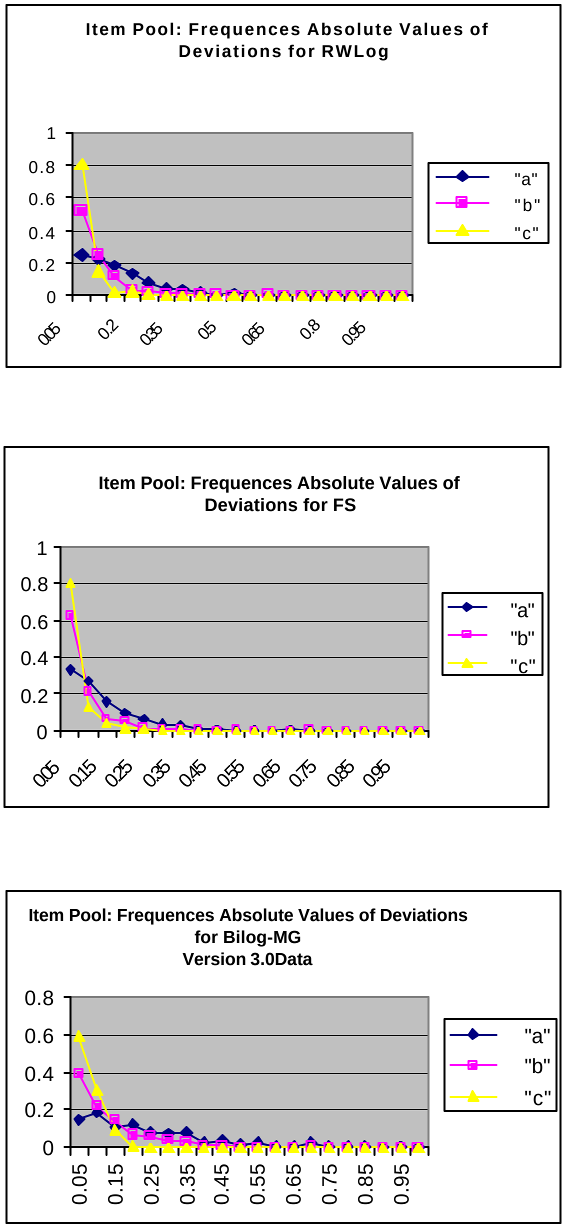


Figure 3. Item Pool Recalibration: Frequencies of deviations of absolute values for three different calibration methods.

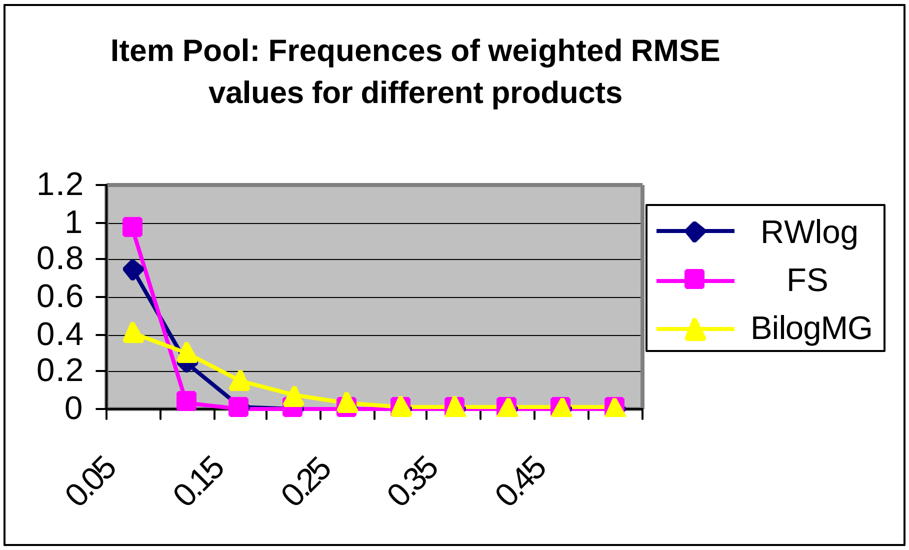


Figure 4. Item Pool Recalibration: Frequencies of RMSE values for three different methods.