Item Calibration/Recalibration in Computerized Multistage Testing (MST) – Evaluation of Some Practical Solutions

Russell Chanjin Zheng

University of Illinois at Urbana-Champagne

September 2012

AICPA Summer Internship Project Report

Item Calibration/Recalibration in Computerized Multistage Testing (MST) – Evaluation of Some Practical Solutions

1. Motivation/Purposes

The accuracy of item parameter estimate is critical in the application of item response theory (IRT) to a large-scale MST testing program, since every aspect of the testing program is based on these estimates, from item selection in the administration, to final ability estimation. In the practice of large scale testing programs, response matrix from the MST administration can be recalibrated to detect the difference between the item parameters obtained from paper-and-pencil (P&P) pretesting or scale drift over time. But the calibration/recalibration of item parameters based on response data from MST administration poses a challenge. Due to the nature of the MST design, the items in one module will be presented to a group of examinees by the routing rule instead of all the examinees. So item response data obtained from MST administration is typically sparse.

One general approach to deal with this issue in MST is to consider it as an equating/linking problem with multiple groups taking multiple forms. There are several advantages for this approach. Firstly, it is conceptually convenient to testing practitioners since equating/linking is an integral part of standard measurement curricula now and an essential element of a psychometrician’s job responsibility. Second, it is convenient to implement since it can be done with existing IRT programs.

There are some potential issues for this approach. For example, in a large-scale MST program, there might be dozens if not hundreds of groups of examinees responding to different test forms. This is rare in regular equating/linking study and calls for study on the feasibility of this approach. The current study intends to investigate the approach more thoroughly, proposing calibration methods to deal with the sparse response matrix, discussing their feasibility in the major IRT programs (BILOG-MG and PARSCALE), and comparing their performance and practical applications in operational testing programs.

2. Perspectives and Theoretical Rationale

2.1 MST Calibration

MST adaptively uses sets of items as the building blocks for a test, analogous to the traditional CAT that adaptively selects individual items for sequential administration to examinees as a test is in progress (Zenisky, Hambleton, & Luecht, 2010). MST can be considered as a special case of CAT since MST is adaptive at item group level and CAT at individual item level. On the other hand, MST has something in common with the P&P fixed form tests. Recently, MST has been proposed as a "balanced compromise" (Hendrickson, 2007) between the traditional paper-and-pencil fixed form tests and CAT.

In the current rising trend of MST, more attention has been paid to the design and delivery of MST than a fundamental issue with significant theoretical and practical importance. Eggen and Verhelst (2011) discussed the justifiability of item parameter estimation in MST. Other than that, there is very little research about the item calibration in MST.

From the discussion of the definition, design and characteristics of MST, there are two possible general approaches to this issue: the first is to consider MST as a special case of CAT and borrow techniques for the calibration and recalibration in CAT; the second is to consider MST as special case of P&P tests and use the equating/linking methods to tackle this issue. The current study is to take the second approach, so the following literature review will be on the calibration and recalibration of CAT response matrix and important equating/linking methods in P&P tests.

2.2 CAT and MST Calibration

Data sparseness and a restricted range of examinee ability has been an important research topic in CAT (Hanson & Béguin, 2002; Haynie & Way, 1995; Hsu, Thompson, & Chen, 1998; Ito & Sykes, 1994; Parshall, 2002; Stocking, 1988).

Several researches demonstrated the negative effect of data sparseness and restricted range of ability on item parameter estimation. Haynie and Way (1995) pointed out the two of the most critical issues in CAT data calibration: data sparseness and restricted range of examinee ability. Ito and Sykes (1994) demonstrated that there was a problem with recalibrating data based on a restricted range of examinee ability. In this Rasch model simulation study, the result indicated that the b values were not well replicated when difficult items were calibrated using responses from able examinees and easy items were calibrated from less able examinees. Hsu et al. (1998) carried out a simulation study that demonstrated the sparseness of CAT administration data had an effect on the precision or accuracy of item recalibration.

Some researches in CAT have been devoted to the theoretical justification or solution to how to address the data sparseness issue in CAT. Mislevy and Wu (1996) have shown that in incomplete designs the justifiability of the marginal maximum likelihood (MML) estimation can be deduced from Rubin and Little's (2002) general theory on inference in the presence of missing data. Ban et al (2001) compared five online calibration methods with different sample sizes in terms of item parameter recovery when there was sparse data only on operational items. They reported that the MML estimate with multiple EM cycles (MEM), which is similar to multiple weights updating or MWU-MEM, is the best method. Ban et al (2002) extended the previous study to a scenario in which the item responses to the pretest items in the pool are sparse. MEM performed the best in that case.

2.3 Equating/linking and MST Calibration

The concurrent calibration (CC) in the linking and equating study can be considered as the special case of the sparse data calibration of the paper-pencil testing. Concurrent calibration, the case in equating studies where items are calibrated from response matrix obtained in incomplete designs, was compared with linking on the same scale separately calibrated tests with data from complete designs (Eggen & Verhelst, 2011; Hanson & Béguin, 2002). In concurrent calibration, response of not-present items were considered as missing data and handled with EM algorithm in many IRT program, such as the concurrent calibration in BILOG-MG.

It is possible to calibrate the response matrix obtained from the MST administration as a concurrent calibration problem. The items that have not been exposed to examinees at Stage 2 and 3 or from other panels can be treated as Not-Present items in the IRT programs. For a practical MST testing program, the missing rate of the response will be high, so the concurrent calibration might potentially introduce more errors.

An alternative linking procedure is called the fixed parameter calibration (FPC) or item anchoring calibration, which combines features of concurrent and separate estimation, is to estimate item parameters for one form and then estimate the parameters in the other form with the common item parameters fixed at their estimated values using the first form (Hanson & Béguin, 2002; Kim, 2006). Kim (2006) compared five fixed parameter calibration (FPC) methods and recommended multiple weights updating and multiple EM cycle (MWU-MEM) which is essentially same as the one proposed by Ban et al (2001). He also pointed out that FPC has the same features as the online calibration except that the latter typically deals with sparse response data because, apart from the “seeded” new-item set, different sets of operational items are adaptively administered to examinees. Kang and Petersen (2009) replicated the result for the FPC and, however, they pointed out that when the two groups had similar ability distributions, FPC by BILOG produced similar result as the concurrent calibration, separate calibration, and FPC by PARSCALE.

It is also possible to apply the FPC to responses from the MST administration. The missing rate for the medium-module items is different from that of the difficulty-module items due to the design of MST in which the medium-module items are exposed to more examinees in Stage 1. All the medium-module items can be considered as a form and all the medium-module and difficult-module another form, so the first form is an inclusive subset of the second form which is different from the traditional FPC. At the same time, only one group of examinees is needed to take the two forms, which is also different form the traditional FPC. The concurrent method can be applied to the response matrix for the medium-module items first and then all the difficult-module items can be calibrated via FPC with all the medium-module items’ parameters fixed.

Software Choice

There are several programs, commercial or noncommercial, available to dealing with the sparse data from a restricted range of ability, specifically, the IRT Command Language (Hanson, 2002), BILOG-MG, PARSCALE, and MULTILOG.

Ban et al (2001) compared ICL and BILOG-MG and found the latter performed worse when dealing the sparse data. Kim (2006) used the ICL, BILOG-MG and PARSCALE. ICL and PARSCALE were able to update the posterior distribution while BILOG-MG was not, so they performed better than the latter. Kang and Petersen (2009) used BILOG-MG and PARSCALE to execute the multiple EM cycles, and the appropriate use of PARSCALE consistently provided better item parameter to the separate and concurrent calibration.

Both BILOG-MG and PARSCALE use multiple EM cycles. But in BILOG-MG, use of the EMPIRICAL command which enables updating of the prior ability distribution overrides use of the NOADJUST command which prevents rescaling of parameters. This problem does not occur with use of PARSCALE, where the NOADUST command which prevents rescaling can be used along with the POSTERIOR command which enables updating of the prior ability distribution multiple times (Kang & Petersen, 2009; Kim, 2006). There is no research pertaining to the performance of MULTILOG in this aspect.

There are several ways to remedy the problems with BILOG-MG with regard to the FPC. Firstly, As Kang and Petersen (2009) pointed out when the two groups have the similar ability distributions, FPC by BILOG-MG produced the similar result as the concurrent calibration, separate calibration and FPC by PARSCALE. Secondly, wise manipulation of BILOG-MG can mitigate the problem, such as the two- run method suggested by Kim (2006) and the one-run method by DeMars and Jurich (2012).

For the current study, BILOG-MG and PARSCALE are chosen. BILOG-MG is the focus of the study firstly because it is the most popular IRT program and the Company is using it for operation. PARSCALE is also one of the most popular IRT programs, although not usually for dichotomous items and it is recommended for the FPC method. The FPC using BILOG-MG in the current study is expected to produce similar result to the concurrent calibration and FPC using PARSCALE because the response matrixes for the FPC are from the same group or groups of the same population.

So based on the previous literature and computer software review, the following calibration methods can be proposed to address the calibration issue in MST:

Method 1: concurrent calibration for the whole response matrix;

Method 2: concurrent calibration for response matrix of all the medium-module items and then using FPC for the difficult-module items (See the simulation design section for the context of this).

Due to different choices of IRT programs, especially for FPC, we have two variants for each of the two methods: Method1A and Method 2A using BILOG-MG, Method 1B and Method 2B using PARSCALE. The advantages and disadvantages are described as follow:

Method	Advantages	Disadvantages
Method 1A/ Method1B	✓ Convenient to implement	✓ May not be able to complete the calibration due to the large percent of missing data
Method 2A/ Method 2B	✓ Easy to implement ✓ Take advantage of the best part of the response matrix	✓ More calibration error might be introduced by discarding the information from the difficult-module items

3. Research Questions

The current study attempts to answer the following questions with regard to the item calibration and recalibration in MST:

First, how well can the response matrix obtained from MST administration be calibrated using the existing IRT programs, especially BILOG-MG?

Second, which calibration method is more suitable for the calibration in MST, concurrent calibration or FPC?

4. Simulation Design

General design

The study uses simulated data to investigate the 4 proposed calibration methods under two different sample sizes. More specifically, the 4 methods are concurrent calibration with BILOG-MG and PARSCALE, FPC with BILOG-MG and PARSCALE. The scenario for using BILOG-MG to calibrate a complete response matrix (no missing data) is also simulated to serve as the baseline. The number of examinees taking a MST panel will be set to 500 or 1000.

The MST design

In the CPA Uniform Exam, the following simplified MST model is adopted which only includes medium modules (M) and difficult modules (D) since the easy modules contribute little to the licensure decision. There are 4 different forms (routes) in every panel, namely, M-M-M, M-M-D, M-D-M, and M-D-D.

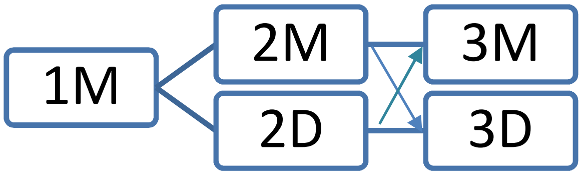

Figure 1 the Simplified MST Model

Item bank generation

Summary statistics (rounded) for a real item bank are obtained and presented as below. 5000 items were generated with the three parameters with the same distributions.

Table 1: Summary Statistics for Item Parameters

	Mean	SD
a	0.75	0.2
b	0.25	1
c	0.25	0.05

Panel construction

24 medium-difficult modules and 16 high-difficult modules in total were constructed from the item bank. In order to make the design simpler, we assume there are no overlapping items among modules and every module is used for 3 times. Thus, 24 (almost) parallel panels were constructed.

Examinee generation

Since there were 24 panels, 24 groups of 500/1000 examinees were randomly selected from N(0,1).

Response generation

Every group of examinees took one panel of items. Within the panel, examinees were routed to different modules at Stage 2 or 3 by the number of items correctly answered in the previous stage(s). The summary statistics of the percentages of examinees taking different routing (forms) across 30 replications for two different group sizes are as follows:

Table 2: Summary Statistics of the Percentages of Different Routings (500 Group Size)

	M-M-M	M-M-D	M-D-M	M-D-D
Mean	40.966	0.151	22.501	36.381
Variance	0.180	0.001	0.087	0.115
Maximum	41.867	0.225	23.067	36.958
Minimum	40.258	0.092	22.008	35.650

Table 3: Summary Statistics of the Percentages of Different Routings (1000 Group Size)

	M-M-M	M-M-D	M-D-M	M-D-D
Mean	40.985	0.146	22.611	36.258
Variance	0.100	0.001	0.119	0.126
Maximum	41.604	0.183	23.296	37.071
Minimum	40.158	0.092	21.833	35.750

Evaluation criteria

This is a 5(4 calibration methods plus one complete matrix calibration) X 2(sample sizes) design. 30 replications were run and item and the MLE ability parameters obtained for every condition. The average bias (from the true values that generated the simulations) was calculated for the item and person parameters (via the EAP algorithm of the program used for calibration). Since the accuracy of examinee classification is one of primary interest to a licensing testing program, the classification accuracy, false negative and false positive rates for the cutoff score 0.5244 normal deviate (the examinee with the ability estimate less than 0.5244 is labeled as non-mastery and the probability of an examinee randomly generated from the standard normal distribution being labeled as non-mastery is 0.7) were also reported for every condition.

5. Result

5.1 **Parameter recovery**

Table 4 presents the average bias (estimates minus true values) of parameter recovery for the 500 group size over 30 replications. MST calibrations produced a slight worse recovery rate than the baseline condition for item parameter, but a better result for the ability parameter. For the concurrent calibration, BILOG-MG produced a better result than PARSCALE. For the FPC, only BILOG-MG can work and PARSCALE cannot perform FPC in the current study which will be discussed in the Discussion section. When BILOG-MG was used, the result was almost identical for the concurrent calibration and FPC. In terms of parameters, the recovery of ability parameter was much better than the item parameters for both methods and IRT programs.

Table 4: Average Bias of Parameter Recovery for the 500 Group Size (30 Replications)

Program	Concurrent Calibration				Fixed parameter Calibration				Complete Matrix (Baseline)			
	a	b	c	ability	a	b	c	ability	a	b	c	ability
Bilog-MG	-0.0745	-0.0709	-0.0316	0.0029	-0.0809	-0.0774	-0.0323	-0.0028	-0.0320	-0.0350	-0.0157	0.0191
Parscale	-0.086	-0.140	-0.0351	0.0005	NA*				NA*			

*: The baseline condition is the complete matrix, so the only the calibration from the BILOG_MG is presented and that of PARSCALE is not necessary.

Table 5 presents the result for the 1000 group size which is very similar to the 500 group size. MST calibrations produced a slight worse recovery rate than the baseline condition for item parameter, but a better result for the ability parameter. The result for PARSCALE is not available. For BILOG-MG, both methods generated similar results as in the 500 group size. The group size effect was not significant and there was a slight improvement in the 1000 group size.

Table 5: Average Bias of Parameter Recovery for the 1000 Group Size (30 Replications)

Program	Concurrent Calibration				Fixed parameter Calibration				Complete Matrix (Baseline)			
	a	b	c	ability	a	b	c	ability	a	b	c	ability
Bilog-MG	-0.0574	-0.0626	-0.0271	0.0013	-0.0609	-0.0720	-0.0279	-0.0063	-0.0208	-0.0246	-0.0106	0.0167
Parscale	NA				NA				NA			

In order to further investigate the difference between the concurrent calibration and FPC, the average biases for the medium-module and difficult-module items were calculated. In the FPC, the medium-module items were calibrated from the response matrix for the medium-module items first and then the difficult-module items were calibrated from the whole matrix with the medium-module items' parameters fixed. Table 6 presents the average biases for the two different types of items from the 500 group size. It indicates that for the two types of items, the bias had the same pattern for the concurrent and FPC, so there was no difference between the concurrent calibration and FPC in the MST item calibration. Table 7 summarizes the bias for the 1000 group size and it indicates the same conclusion.

Table 6: Average Bias for the Medium Modules and Difficulty Modules (500 Group Size)

Items Calibrated	Concurrent Calibration			Fixed Parameter Calibration			Complete Matrix (Baseline)		
	a	b	c	a	b	c	a	b	c
Medium-module	-0.0755	-0.1030	-0.0405	-0.0745	-0.1139	-0.0406	-0.0326	-0.0591	-0.0226
Difficult-module	-0.0735	-0.0492	-0.0256	-0.0852	-0.0522	-0.0267	-0.0318	-0.0196	-0.0112
All items	-0.0745	-0.0709	-0.0316	-0.0809	-0.0774	-0.0323	-0.0320	-0.0350	-0.0157

Table 7: Average Bias for the Medium Modules and Difficulty Modules (1000 Group Size)

Items Calibrated	Concurrent Calibration			Fixed Parameter Calibration			Complete Matrix (Baseline)		
	a	b	c	a	b	c	a	b	c
Medium-module	-0.0545	-0.0850	-0.0330	-0.0535	-0.0969	-0.0334	-0.0212	-0.0404	-0.0153
Difficult-module	-0.0592	-0.0475	-0.0232	-0.0660	-0.0549	-0.0243	-0.0206	-0.0142	-0.0075
All items	-0.0574	-0.0626	-0.0271	-0.0609	-0.0720	-0.0279	-0.0208	-0.0246	-0.0106

5.2 Classification Accuracy

For licensing and credentialing exams, the bias of the parameter recovery is of second importance. A more important issue is the rate of correct binary classification of examinees. Table 8 summarizes the classification accuracy rate for the two methods in BILOG-MG. The correct classification rates are satisfactory. The false negative rate is about 9% which means 9 percent of the examinees who master might be categorized as disqualified. The false positive rate is about 0.5% and this indicates that the probability of awarding the certificate to an incompetent individual is close to zero percentage. The results did not imply method effect or group size effect.

Table 8: Classification Rate (%) for Ability Estimates

	Correct Classification	False Negative	False Positive
500			
BILOG-Concurrent	90.4	9.1	0.5
BILOG-FPC	90.4	9.2	0.4
BILOG-Complete	88.9	11.1	0.0
1000			
BILOG-Concurrent	90.6	8.9	0.5
BILOG-FPC	90.7	8.7	0.5
BILOG-Complete	88.9	11.1	0.0

6. Discussion and Conclusion

The current research demonstrates that there is a possibility of calibrating or recalibrating item parameters from the response matrix of the MST administration. BILOG-MG and PARSCALE produced acceptable item parameter recovery with concurrent calibration, compared to the baseline condition. The BILOG-MG generates the similar result for the FPC and for the concurrent calibration.

The proposed PFC method using PARSCALE is not feasible for the MST item calibration because there is a practical problem to read the item parameter estimates from the first run. In the proposed method, the lengths of the two forms for the two runs are different. In traditional FPC, two forms are usually of the same test length, so the item parameter estimates can be read for the second run (the fixed parameter estimate run) by the command IPNAME which can only read the item parameter output file from the PARSCALES. If the lengths of two forms are different, the program warns that the item numbers are not compatible and refuses to read the item parameter estimate. Therefore the proposed FPC method using the PARSCALE is impractical in MST calibration.

It’s more encouraging that the recovery of the ability parameter is much better than that of the item parameter which echoes the result of Chuah, Drasgow, and Luecht (2006). The classification rate shows that the ability parameter estimates from the response matrix obtained from the MST administration are satisfactory and even better than those of the baseline condition.

Item calibration/recalibration is one of the basic elements of MST testing program. However, there has been a lack of published studies to address the challenges of item calibration in an MST setting. Even in CAT, only some research discussed the negative effects associated with this issue of sparse matrix and necessity of addressing it, and only one study (Harmes, Parshall, & Kromrey, 2003) provided potential solution in need of further validation. In the current rising trend of MST, more attention has been paid to the design and delivery of MST than this fundamental issue with significant theoretical and practical importance. The current study suggests and evaluates a potential solution to the calibration in MST from the perspective of equating/linking that is conceptually easy and convenient to implement. This study contributes to the understanding of the calibration in MST, sheds lights on the connection between the calibration issue in MST and P&P testing, and provides some practical help to the testing practitioners involved in the MST application.

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*(3), 191-212.

Ban, J. C., Hanson, B. A., Yi, Q., & Harris, D. J. (2002). Data Sparseness and On‐Line Pretest Item Calibration‐Scaling Methods in CAT. *Journal of Educational Measurement, 39*(3), 207-218.

Chuah, S. C., Drasgow, F., & Luecht, R. (2006). How big is big enough? Sample size requirements for CAST item parameter estimation. *Applied Measurement in Education, 19*(3), 241-255.

DeMars, C. E., & Jurich, D. P. (2012). Software Note Using BILOG for Fixed-Anchor Item Calibration. *Applied Psychological Measurement, 36*(3), 232-236.

Eggen, T. J. H. M., & Verhelst, N. D. (2011). Item calibration in incomplete testing designs. *Psicológica*(1), 107-132.

Hanson, B. (2002). IRT Command Language (ICL).

Hanson, B. A., & Béguin, A. A. (2002). Obtaining a common scale for item response theory item parameters using separate versus concurrent estimation in the common-item equating design. *Applied Psychological Measurement, 26*(1), 3-24.

Harmes, J. C., Parshall, C. G., & Kromrey, J. D. (2003). *Recalibration of IRT item parameters in a CAT: sparse data matrices and missing data treatments.* Paper presented at the the Annual Meeting of the National Council on Measurement in Education, Chicago.

Haynie, K., & Way, W. (1995). *An investigation of item calibration procedures for a computerized licensure examination.* Paper presented at the the Annual Meeting of the National Council on Measurement in Education, San Fracisco.

Hendrickson, A. (2007). An NCME instructional module on multistage testing. *Educational Measurement: Issues and Practice, 26*(2), 44-52.

Hsu, Y., Thompson, T., & Chen, W. (1998). *CAT item calibration.* Paper presented at the the Annual Meeting of the National Council on Measurement in Education, San Diego.

Ito, K., & Sykes, R. C. (1994). *The Effect of Restricting Ability Distributions in the Estimation of Item Difficulties: Implications for a CAT Implementation*. Paper presented at the the Annual Meeting of the National Council on Measurement in Education.

Kang, T., & Petersen, N. S. (2009). *Linking item parameters to a base scale* (No. 2009-2): ACT.

Kim, S. (2006). A comparative study of IRT fixed parameter calibration methods. *Journal of Educational Measurement, 43*(4), 355-381.

Misley, R. J., & Wu, P. K. (1996). *Missing Responses and IRT Ability Estimation: Omits, Choice, Time Limits, and Adaptive Testing*: ETS.

Parshall, C. G. (2002). Item development and pretesting in a CBT environment. In C. N. Mills, M. Potenza, J. J. Fremer & W. C. Ward (Eds.), *Computer-based testing: building the foundation for future assessments* (pp. 119-141). Mahwah, New Jersy: Lawrence Erlbaum Associates, Publishers.

Rubin, D. B., & Little, R. J. A. (2002). Statistical analysis with missing data. *Hoboken, NJ: J Wiley & Sons*.

Stocking, M. L. (1988). *Scale drift in on-line calibration*: ETS.

Zenisky, A., Hambleton, R. K., & Luecht, R. M. (2010). Multistage testing: Issues, designs, and research. *Elements of Adaptive Testing*, 355-372.