MWSUG 2016 – Paper AA27

Fixed Item Parameter Calibration with MMLE-EM Using a Fixed Prior in SAS/IML®

Sung-Hyuck Lee, ACT, Iowa City, IA
Hongwook Suh, ACT, Iowa City, IA

ABSTRACT

Fixed item parameter calibration (FIPC) has been popular in calibrating the parameters of pretest (new) items administered with a computer adaptive test. In this study, a new FIPC method is proposed. In the new approach, the prior for the EM algorithm is computed using only the parameters of operational (scored) items and the responses to these items for the new calibration sample. During the EM cycles the prior is not updated but fixed in calibrating pretest items. The main advantage of this new method is that any potential contamination from poor pretest items (e.g., poor model fit) is eliminated since pretest items are excluded from computing the prior during the EM cycles. No commercial software is available to implement the new approach so a new SAS® macro named SAS®-FIPC is written in SAS/IML® to calibrate the parameters of the pretest items. The calibration results of the new method are compared to the ones from the existing methods through a simulation study.

INTRODUCTION

Estimating item parameters is one of the most fundamental procedures to maintain the integrity of test scores from the item response model framework (Baker & Kim, 2004). In practice, when the parameters of pretest (new) items are estimated, their parameter estimates must be put on the same scale as operational (scored) items that are already on the base scale, by means of a scale transformation of the estimated pretest items to the base scale. It is an unavoidable process for the pretest items to be used as scored items in later administrations. Only when item parameters are placed on the same scale, test scores computed based on their item parameters are comparable to each other.

The fixed item parameter calibration (FIPC) has been popular for online tests (Paek & Young, 2005). It allows for test practitioners to estimate the parameters of pretest items that are embedded among operational items and place them on the same sale as the operational items. Unlike separate estimations of operational and pretest items which are linked together with a scale transformation (e.g., Stocking & Lord, 1983), the FIPC method calibrates pretest items with the parameters of operational items fixed at the values obtained previously. Therefore, there is no scale indeterminacy issue with the FIPC method.

The success of the FIPC depends on how well the underlying ability distribution is estimated in the current group relative to the base group and on how to use the prior density during the calibration process. The Marginal Maximum Likelihood Estimation method with EM algorithm (MMLE-EM), most commonly employed for the FIPC, estimates the underlying ability distribution (e.g., posterior) in the E-step and use it as the prior in the M-step of the EM cycles. In previous studies (e.g., Ban et. al., 2001; Kim, 2006; Wainer & Mislevy, 2000), various FIPC methods are classified according to whether the prior for the MMLE-EM is updated and how many times it is updated during the EM cycles. As the prior density is properly updated through the EM cycles, it approaches the underlying ability distribution relative to the base scale, and the estimated parameters of pretest items are placed onto the scale of the operational items (i.e., the base scale).

Since the existing FIPC uses not only operational items but also pretest items in updating the prior during the EM cycles, it is possible that the underlying density is contaminated when pretest items show poor model fit. In reality, many newly developed items are discarded or revised due to their bad model fit. Unfortunately, it has not yet been addressed in the literature how a poorly estimated prior density affects the calibration results of the FIPC method and how it can be resolved.

In this study, a new FIPC method is proposed, in which the underlying ability distribution is estimated prior to pretest item calibration according to Mislevy (1984) using only the operational item parameters and the responses to the operational items for the current group of examinees used in the calibration, and is fixed during the EM cycles. The main advantage of this new method is that any potential contamination from bad pretest items (e.g., poor model fit) is eliminated since pretest items are excluded from computing the prior density during the EM cycles. So, it is expected that the new FIPC method, in which pretest items with poor model fit are excluded from the calibration of the (fixed) underlying density, performs better than existing FIPC methods.

METHOD

No commercial software is available to implement the new approach so a new SAS® macro named SAS®-FIPC is written in SAS/IML® to calibrate the parameters of the pretest items. A simulation study was conducted to evaluate the accuracy of SAS®-FIPC in recovering the item parameters. In this study, three simulation factors were manipulated: the number of examinees, the number of pretest items with poor model fit, and the underlying ability distribution. For the number of examinees, 1,000 and 2,000 examinees were chosen to represent medium and large testing volumes. A 50-item test was simulated with 10 operational items and 40 pretest items. Among the 40 pretest items, 0%, 25%, 50%, or 75% of items were assumed to have poor model fit. Therefore, including the no model misfit condition, four model fit conditions in the pretest items were examined. The following 5 underlying ability distributions were examined: $N(0, 1^2)$,$N(0.5, 1.2^2)$, $N(-0.5, 1.2^2)$, $N(1, 1.2^2)$, and $N(-1, 1.2^2)$.

The three-parameter logistic model (3PLM) was used when generating item responses. However, pretest items with poor model fit are simulated by generating examinee responses using the one-parameter logistic model (1PLM). 80 conditions were simulated, in which the 2 X 4 X 5 simulation factors were crossed with two different calibration programs: SAS®-FIPC for the new FIPC method (written in SAS/IML®) and PARSCALE (Muraki & Bock, 1998). The fixed prior (e.g., underlying ability distribution) was estimated using the procedure suggested by Mislevy (1984) and the fixed prior was not updated during the EM cycles. One hundred replications were simulated for each of the 80 conditions.

GENERATING SIMULATED DATA RESPONSES

The ability parameters of examinees were generated from five sets of normal distributions as mentioned previously. To address various ranges of underlying ability distributions, means and standard deviations of the distributions were modified within the level of acceptable differentiations. Table 1 shows the descriptive statistics of simulated examinee's true ability (θ) distributions for the two sample size conditions.

Table 1. Descriptive statistics for simulated examinee ability parameter

Condition	N	Mean	SD
$N(0, 1^2)$	1000	0.022	0.958
	2000	-0.010	1.019
$N(0.5, 1.2^2)$	1000	0.521	1.120
	2000	0.498	1.100
$N(-0.5, 1.2^2)$	1000	-0.445	1.113
	2000	-0.494	1.129
$N(1, 1.2^2)$	1000	1.093	1.180
	2000	1.008	1.178
$N(-1, 1.2^2)$	1000	-1.012	1.259
	2000	-0.984	1.168

Item parameters were obtained from a large scale standardized mathematics test. Table 2 shows the summary statistics for the item parameters used to generate examinee responses. The parameters of operational items show a slightly higher item discrimination (*a*) and lower item difficulty (*b*) values than those for pretest items.

Table 2. Descriptive statistics for item parameters

Item	Item Parameter Mean (SD)		
	a	*b*	*c*
Operational items	0.910 (0.164)	-0.128 (0.556)	0.227 (0.080)
Pretest items	0.766 (0.294)	0.004 (1.145)	0.198 (0.110)
Overall	0.730 (0.309)	0.037 (1.253)	0.191 (0.116)

EVALUATION OF THE OUTCOMES

The results from SAS®-FIPC are compared with those from a commonly used commercial IRT program, PARSCALE. Similar calibration parameters were implemented for both programs (e.g., maximum EM cycles and the number Newton-Raphson iteration). Outcomes for the two different calibration programs were compared by calculating the Mean Bias Error (MBE) and the Root Mean Squared Error (RMSE), calculated as follows:

$$\text{MBE} = \frac{\sum_{i=1}^{n}(X_i - X_t)}{n},$$

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(X_i - X_t)^2}{n}},$$

where X_t is the true parameter, X_i is the estimator of the corresponding parameter, and *n* is the number of replications (= 100). Each parameter estimate was compared with the corresponding true parameter value by calculating the deviation and squared deviation, then averaged over replications in computing MBE and RMSE, respectively.

RESULTS

It was noticed that the number of successful convergences (across replications) in both programs were different depending on the conditions. The number of replications with non-convergence results increased for both programs as the underlying ability distribution deviated from the $N(0, 1^2)$ condition and as the number of pretest items with poor fit increased. However, there was no consistent difference in number of non-convergence results between the two estimation programs. Only replications that showed successful convergence were included when calculating evaluation criteria.

The MBE and RMSE between the true and estimated parameters were averaged over the replications to evaluate the performance of SAS®-FIPC and PARSCALE in the recovery of the true parameters. Table 3 shows that when no pretest items show poor model fit, SAS®-FIPC generally shows lower MBE and RMSE than PARSCALE in recovering the *a*, *b*, and *c* parameters (although *b* shows lower mean bias for PARSCALE in conditions with low mean ability). As expected, RMSE is lower in the N = 2000 condition than in the N = 1000 condition for all item parameters. However, although MBE is smaller in the N = 2000 condition for *a*, it is larger in the N = 2000 condition for *b* and *c*, especially in the sampling distribution conditions with high negative mean. Also, the *a* and *b* parameters appear to have lower RMSE in sampling distribution conditions with higher positive mean, and higher RMSE in sampling distribution conditions with higher negative mean, especially when the parameters are estimated by PARSCALE.

Table 3. Mean Bias and Mean RMSE when no pretest items with poor fit are present

Condition	N	Mean Bias						Mean RMSE					
		SAS®-FIPC			PARSCALE			SAS®-FIPC			PARSCALE		
		a	*b*	*c*	*a*	*b*	*c*	*a*	*b*	*c*	*a*	*b*	*c*
$N(0, 1^2)$	1000	**-0.030**	**-0.059**	**-0.016**	-0.096	0.061	-0.031	**0.160**	**0.342**	**0.095**	0.226	0.486	0.100
	2000	**-0.022**	-0.085	**-0.027**	-0.064	**0.047**	-0.042	**0.121**	**0.310**	**0.088**	0.176	0.476	0.096
$N(0.5, 1.2^2)$	1000	**-0.016**	**-0.060**	**-0.006**	-0.087	-0.064	-0.024	**0.126**	**0.326**	**0.093**	0.165	0.385	0.097
	2000	**-0.004**	**-0.063**	**-0.013**	-0.057	-0.091	-0.029	**0.101**	**0.285**	**0.086**	0.120	0.322	0.089
$N(-0.5, 1.2^2)$	1000	**-0.056**	-0.072	**-0.031**	-0.098	**0.028**	-0.038	**0.173**	**0.338**	**0.090**	0.219	0.456	0.093
	2000	**-0.042**	-0.086	**-0.033**	-0.065	**0.036**	-0.045	**0.140**	**0.295**	**0.081**	0.181	0.447	0.088
$N(1, 1.2^2)$	1000	**-0.061**	**-0.065**	**0.005**	-0.076	-0.109	-0.018	**0.131**	**0.335**	**0.098**	0.146	0.375	0.102
	2000	**0.001**	**-0.054**	**-0.006**	-0.052	-0.110	-0.025	**0.091**	**0.291**	**0.090**	0.108	0.324	0.094
$N(-1, 1.2^2)$	1000	**-0.075**	-0.049	**-0.029**	-0.115	**0.029**	-0.034	**0.196**	**0.344**	**0.079**	0.232	0.425	0.081
	2000	**-0.053**	-0.075	**0.032**	-0.067	**0.069**	-0.043	**0.160**	**0.320**	**0.075**	0.202	0.454	0.082

* *Note:* Bold typed number indicates smaller value in absolute number.

Table 4 shows similar results when 25% of pretest items show poor model fit, in which SAS®-FIPC generally show better recovery than PARSCALE across conditions. The RMSE values are generally smaller in Table 4 than in Table 3 (no items with poor fit) for *a* and *b* parameters estimated by SAS®-FIPC, but are larger in Table 4 than in Table 3 for *a*, *b*, and *c* parameters estimated by PARSCALE, especially for the *b* parameter. Furthermore, the expected decrease in RMSE with sample size breaks down when *b* and *c* parameters are estimated by PARSCALE. Also, the effect of sampling distribution condition on RMSE appears diminished in Table 4, except for the *b* parameter when it is estimated by PARSCALE.

Table 4. Mean Bias and Mean RMSE when 25% of pretest items show poor model fit

Condition	N	Mean Bias						Mean RMSE					
		SAS®-FIPC			PARSCALE			SAS®-FIPC			PARSCALE		
		a	*b*	*c*	*a*	*b*	*c*	*a*	*b*	*c*	*a*	*b*	*c*
$N(0, 1^2)$	1000	**-0.023**	**0.001**	**-0.005**	-0.166	0.221	-0.018	**0.151**	**0.270**	**0.086**	0.338	0.627	0.127
	2000	**-0.020**	**-0.012**	**-0.012**	-0.147	0.225	-0.035	**0.116**	**0.207**	**0.072**	0.300	0.641	0.136
$N(0.5, 1.2^2)$	1000	**-0.009**	**-0.048**	**-0.013**	-0.189	0.083	-0.007	**0.117**	**0.315**	**0.095**	0.342	0.502	0.122
	2000	**-0.009**	**-0.065**	**-0.021**	-0.178	0.102	-0.014	**0.093**	**0.278**	**0.090**	0.315	0.515	0.126
$N(-0.5, 1.2^2)$	1000	**-0.053**	**-0.082**	**-0.036**	-0.153	0.210	-0.020	**0.164**	**0.310**	**0.093**	0.334	0.608	0.132
	2000	**-0.038**	**-0.079**	**-0.035**	-0.088	0.240	-0.043	**0.122**	**0.260**	**0.080**	0.257	0.685	0.147
$N(1, 1.2^2)$	1000	**-0.016**	**-0.054**	**0.001**	-0.157	-0.116	0.005	**0.107**	0.342	**0.103**	0.322	**0.331**	0.119
	2000	**0.005**	**-0.047**	**-0.012**	-0.122	-0.120	0.002	**0.084**	**0.281**	**0.092**	0.282	0.283	0.119
$N(-1, 1.2^2)$	1000	**-0.064**	**-0.079**	**-0.031**	-0.166	0.155	-0.005	**0.189**	**0.318**	**0.078**	0.352	0.507	0.126
	2000	**-0.054**	**-0.080**	**-0.032**	-0.049	0.310	-0.029	**0.166**	**0.302**	**0.074**	0.262	0.687	0.135

* *Note:* Bold typed number indicates smaller value in absolute number.

Tables 5 and 6 show a similar pattern when 50% and 75% of pretest items show poor model fit. Again, SAS®-FIPC shows consistently better item parameter recovery across all of the conditions. Also, RMSE appears to decrease as the percentage of poor fit items increases when the item parameters are estimated by SAS®-FIPC, but increases with percentage of poor fit items when the item parameters are estimated by PARSCALE. In addition, RMSE is smaller in the N = 2000 condition than in the N = 1000 conditions when the *a*, *b*, and *c* parameters are estimated by SAS®-FIPC, but is larger in the N = 2000 condition than in the N = 1000 condition when *b* and *c* parameters are estimated by PARSCALE (although RMSE is still smaller in the N = 2000 condition when the *a* parameter is estimated by PARSCALE). Finally, the *b* and *c* parameters appear to have lower RMSE in sampling distribution conditions with higher positive mean, and higher RMSE in sampling distribution conditions with higher negative mean, when the parameters are estimated by PARSCALE, but the effect is reversed when parameters are estimated by SAS®-FIPC. However, the *a* parameters appear to have *higher* RMSE in sampling distribution conditions with higher positive mean, and lower RMSE in sampling distribution conditions with higher negative mean, when the parameters are estimated by PARSCALE, but the effect is reversed when parameters are estimated by SAS®-FIPC.

Table 5. Mean bias and RMSE when 50% of pretest items show poor model fit

Condition	N	Mean Bias						Mean RMSE					
		SAS®-FIPC			PARSCALE			SAS®-FIPC			PARSCALE		
		a	*b*	*c*	*a*	*b*	*c*	*a*	*b*	*c*	*a*	*b*	*c*
$N(0, 1^2)$	1000	**-0.017**	**0.005**	**-0.006**	-0.260	0.206	0.005	**0.147**	**0.252**	**0.082**	0.435	0.624	0.144
	2000	**-0.023**	**-0.024**	**-0.017**	-0.221	0.209	-0.006	**0.109**	**0.200**	**0.070**	0.380	0.648	0.155
$N(0.5, 1.2^2)$	1000	**-0.006**	**0.025**	**0.004**	-0.293	0.135	0.015	**0.127**	**0.251**	**0.083**	0.442	0.481	0.138
	2000	**0.002**	**-0.003**	**-0.005**	-0.268	0.107	0.008	**0.095**	**0.211**	**0.072**	0.414	0.596	0.145
$N(-0.5, 1.2^2)$	1000	**-0.051**	**-0.019**	**-0.019**	-0.222	0.181	0.005	**0.161**	**0.223**	**0.068**	0.396	0.733	0.150
	2000	**-0.044**	**-0.028**	**-0.018**	-0.179	0.116	0.002	**0.127**	**0.183**	**0.057**	0.356	0.836	0.155
$N(1, 1.2^2)$	1000	**0.001**	**0.035**	**0.018**	-0.266	-0.083	0.021	**0.108**	**0.272**	**0.094**	0.427	0.252	0.129
	2000	**0.004**	**0.007**	**0.002**	-0.208	-0.152	0.022	**0.085**	**0.225**	**0.081**	0.378	0.417	0.139
$N(-1, 1.2^2)$	1000	**-0.058**	**-0.009**	**-0.012**	-0.237	0.187	0.031	**0.177**	**0.208**	**0.054**	0.410	0.606	0.144
	2000	**-0.052**	**-0.025**	**-0.016**	-0.142	0.199	0.017	**0.145**	**0.171**	**0.046**	0.335	0.807	0.148

* *Note:* Bold typed number indicates smaller value in absolute number.

Table 6. Mean bias and RMSE when 75% of pretest items show poor model fit

Condition	N	Mean Bias						Mean RMSE					
		SAS®-FIPC			PARSCALE			SAS®-FIPC			PARSCALE		
		a	*b*	*c*	*a*	*b*	*c*	*a*	*b*	*c*	*a*	*b*	*c*
$N(0, 1^2)$	1000	**-0.028**	**-0.012**	**-0.019**	-0.332	0.182	0.038	**0.144**	**0.215**	**0.073**	0.480	0.597	0.166
	2000	**-0.041**	**-0.053**	**-0.028**	-0.277	0.203	0.035	**0.110**	**0.181**	**0.066**	0.423	0.690	0.178
$N(0.5, 1.2^2)$	1000	**-0.008**	**-0.011**	**-0.013**	-0.356	0.096	0.044	**0.115**	**0.231**	**0.072**	0.485	0.466	0.159
	2000	**-0.017**	**-0.019**	**-0.016**	-0.317	0.024	0.044	**0.093**	**0.190**	**0.066**	0.454	0.614	0.169
$N(-0.5, 1.2^2)$	1000	**-0.074**	**-0.048**	**-0.026**	-0.296	0.151	0.052	**0.165**	**0.197**	**0.063**	0.450	0.717	0.175
	2000	**-0.059**	**-0.046**	**-0.024**	-0.239	0.122	0.055	**0.125**	**0.158**	**0.051**	0.406	0.802	0.178
$N(1, 1.2^2)$	1000	**-0.005**	**0.008**	**0.003**	-0.335	-0.109	0.043	**0.104**	**0.242**	**0.079**	0.466	0.237	0.144
	2000	**-0.004**	**-0.008**	**-0.012**	-0.246	-0.236	0.048	**0.084**	**0.207**	**0.071**	0.404	0.551	0.163
$N(-1, 1.2^2)$	1000	**-0.071**	**-0.029**	**-0.014**	-0.317	0.155	0.081	**0.161**	**0.168**	**0.045**	0.463	0.593	0.175
	2000	**-0.067**	**-0.025**	**-0.018**	-0.217	0.148	0.070	**0.141**	**0.142**	**0.041**	0.399	0.829	0.178

* *Note:* Bold typed number indicates smaller value in absolute number.

Overall both SAS®-FIPC and PARASCALE estimate parameters of pretest items quite well regardless of the number of examinees and the number of items with poor fit. Mean bias and RMSE values did not exceed more than 1 in any of the conditions. However, SAS®-FIPC showed consistently better performance in most of the simulation conditions, although the difference between the two estimation programs are relatively negligible in most conditions. Further research under more various conditions is recommended so that the advantages and disadvantages of the two estimation programs are revealed.

REFERENCE

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

Baker, F. B., & Kim, S.-H. (2004). *Item response theory: Parameter estimation techniques* (2nd ed.). New York: Marcel Dekker.

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

Mislevy, R. J. (1984). Estimating latent distributions. *Psychometrika, 49,* 359–381.

Muraki, E., & Bock, R. D. (1998). *PARSCALE (Version 3.5): IRT item analysis and test scoring for rating-scale data[Computer program]*. Lincolnwood IL: Scientific Software.

Paek, I., & Young, M. J. (2005). Investigation of student growth recovery in a fixed-item linking procedure with a fixed-person prior distribution for mixed-format test data. *Applied Measurement in Education, 18,* 199–215.

Stocking, M., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement*, *7*, 207-210.

Wainer, H., & Mislevy, R. J. (2000). Item response theory, item calibration, and proficiency estimation. In H. Wainer (Ed.), *Computerized adaptive testing*: A primer (2nd ed., pp. 85-87), Mahwah, NJ: Lawrence Erlbaum.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Sung-Hyuck Lee
ACT, Inc
319-341-2417
Sung.lee@act.org

Hongwook Suh
ACT, Inc
319-341-2257
Hongwook.Suh@act.org