The Factor Analytic Method for Item Calibration under Item Response Theory: A Comparison Study Using Simulated Data

Arai, Sayaka
The National Center for University Entrance Examinations, Research Division
2-19-23 Komaba, Meguro-ku
Tokyo 153-8501, Japan
E-mail: sayarai@rd.dnc.ac.jp

Mayekawa, Shin-ichi
Tokyo Institute of Technology, Graduate School of Decision Science and Technology
2-12-1 Ookayama, Meguro-ku,
Tokyo 152-8550, Japan
E-mail: mayekawa@hum.titech.ac.jp

Introduction

Latent trait theory is often referred to as item response theory (IRT) in the area of educational testing and psychological measurement. IRT models show the relationship between the unobserved constructs (e.g., an academic proficiency) and the observed variables (e.g., an item response of the examinee). Because IRT provides many advantages over classical test theory, IRT methods are used in many testing applications. One of the useful features of IRT is the comparability of the test scores obtained from different test forms. However, the parameters of test items need to be put onto the common metric, namely the item parameter calibration, in advance.

Among various IRT models, this study focuses on unidimensional IRT models for dichotomously (0/1) scored tests. Under the three-parameter logistic (3PL) model (Lord, 1980), the probability of a correct response to the item j for the latent trait variable θ is defined as

$$P(\theta|a_j, b_j, c_j) = c_j + \frac{1 - c_j}{1 + \exp(-1.7a_j(\theta - b_j))},$$

where a_j, b_j, and c_j are item parameters for the item j and are alos referred to as the discrimination parameter, difficulty parameter, and pseudo-guessing parameter for item j, respectively. Under the two-parameter logistic (2PL) model, the value of the pseudo-guessing parameter is fixed at zero. Thus, the model is defined as

$$P(\theta|a_j, b_j) = \frac{1}{1 + \exp(-1.7a_j(\theta - b_j))}.$$

Let us assume that θ is linearly transformed by $\theta^* = u + v\theta$, where u and v are the scale transformation constants. The linear transformation of the item parameters, $a_j^* = a_j/v$ and $b_j^* = u + vb_j$, would produce the same probability of a correct response for the item j. This property indicates that the parameter estimates on different IRT scales are linearly related. Therefore, a proper linear transformation allows the parameter estimates on different scales to be converted onto the same scale (Kolen & Brennan, 2004).

Several item calibration methods are used for linking item parameters to the same scale. The calibration methods are generally classified into three groups: (a) separate calibration, (b) concurrent calibration, and (c) fixed common item parameter calibration. The performances of these methods have been compared in many studies (Hanson & Béguin, 2002; Kim & Lee, 2006; Lee & Ban, 2010).

The factor analytic method, which was first proposed by Mayekawa (1991), had not been examined until recently. This method is categorized as one of the characteristic curve methods for separate

calibration. The first study (Arai & Mayekawa, 2011) using the factor analytic method showed that it performed well, but they did not compare it to the other item characteristic curve methods . Another study showed that it performed best among the other methods, but because they had used practical data, we could not assess the accuracy (Fujita & Mayekawa, to appear).

In this study, we focused on five calibration methods: the factor analytic method, the two characteristic curve methods (Stocking-Lord and Haebara), the moment method (Mean/Sigma), and the fixed common item parameter (FCIP) calibration methods. Characteristic curve methods and moment methods are subcategories of separate calibration methods. Simulated data were generated assuming two IRT models. Using various simulation conditions, we examined the relative performance and robustness of these calibration methods.

Calibration methods

In separate calibration, item parameters for two test forms are first separately estimated. Next, the two sets of item parameter estimates of the common items are used to estimate the scale transformation constants, u and v, used for linking the two scales.

The Mean/Sigma method is a moment methods. It uses the means and standard deviations of the b-parameter estimates of the common items (Marco, 1977). The Haebara method and the Stocking-Lord method both use the characteristic curves of common items. The Haebara method finds scale transformation constants such that the sum of the squared differences between the item characteristic curves are minimized (Haebara, 1980). The Stocking-Lord method finds constants such that the squared difference between the test characteristic curves are minimized (Stocking & Lord, 1983).

The factor analytic method minimizes the criterion

$$\Sigma_{g=1}^{G}\Sigma_{j\in g}\int_{\Theta}[P(\theta|\hat{a}_j^{(g)},\hat{b}_j^{(g)},\hat{c}_j^{(g)})-P(-\frac{u_g}{v_g}+\frac{1}{v_g}\theta|a_j,b_j,c_j)]^2h_g(\theta)d\theta,$$

where $\hat{a}_j^{(g)}$, $\hat{b}_j^{(g)}$, and $\hat{c}_j^{(g)}$ are item parameter estimates separately calibrated for the form g, $g = 1, 2, ..., G$; u_g, and v_g are the scale transformation constants for the form g; and $h_g(\theta)$ is the ability distribution of the form g (Arai & Mayekawa, 2011).

In the FCIP method, item parameters are estimated using two separate runs of the estimation program. Unlike separate calibrations, in the second calibration run, item parameters for the common items are fixed at the values estimated in the first calibration run.

Methods

Simulations were designed based on the study of Hanson and Béguin (2002). We used two test forms, the old form (Form A) and the new form (Form B). Both test forms consisted of 60 items, and they had 20 common items. In this study, the item parameters for Form B were to be put on the scale of Form A.

One hundred sets of item parameters, a, b, and c, were generated for the 3PL model such that $a \sim$ LN(0, 0.2), $b \sim$ N(0, 1), and $c \sim$ BETA(8, 32). Another 100 sets of item parameters, a and b were generated for the 2PL model from the same distributions such that $a \sim$ LN(0, 0.2) and $b \sim$ N(0, 1). These item parameters were divided into five sets of 20 items such that the statistical characteristics of five sets were as similar as possible. Form A consisted of the first three sets of 20 items, i.e., Item 1 through Item 60. Form B consisted of the first set and the last two sets of 20 items, i.e., Item 1 through Item 20 and Item 61 through Item 100.

The proficiency variables (θ) for Form A were sampled from a normal distribution with mean 0 and standard deviation 1; this was denoted as N(0, 1). Three sets of the proficiency variables for

Form B were sampled from the N(0, 1), N(0.5, 1), and N(1, 1) distributions. One hundred sets of these four proficiency variables were drawn and then item response data sets were generated using WinGen3 (Han & Hambleton, 2010).

Two levels of sample sizes were considered: 500 examinees per form and 3000 examinees per form. Three levels of the number of common items were considered: 5 items, 10 items, and 20 items. The 20 common items were divided into four sets of five items such that the statistical characteristics of the four sets were as similar as possible. The first set of five items were used as common items in the 5-common item condition, and the first two sets was used as common items in the 10-common item condition. The rest of the 15 (or 10) items were treated as unique (non-common) items.

Four factors were considered for this simulation study: IRT models (2PL and 3PL); proficiency distribution levels (N(0, 1), N(0.5, 1), and N(1, 1)); sample sizes (500 and 3,000); and the number of common items (5, 10, and 20). There were a total of 36 conditions. Under each condition, the five calibration methods were compared; factor analytic (FA) method, Stocking-Lord (SL) method, Haebara (HB) methods, Mean/Sigma (MS) method, and FCIP (FI) method.

Item parameters for the two forms (Form A and Form B) were separately estimated using the computer program BILOG-MG (Zimowski, Muraki, Mislevy, & Bock, 2003). To estimate the scale transformation constants, a program called CALR was used for the FA method and an R package "plink" was used for the SL, HB, and MS methods. In the FI method, item parameters estimated by BILOG-MG for Form B were not used, but the computer program PARSCALE (Muraki & Bock, 2003) was used for the estimation of item parameters for Form B. PARSCALE was run twice. The initial values for the unique items were estimated in the first run, and the item parameters on the scale of Form A were estimated in the second run. This procedure is based on the revised FCIP method (Kim, 2006; Kang & Petersen, 2009).

The criterion used in this study is the average of the mean squared error (MSE) of the item characteristic curves (ICCs). This ICC criterion is given by

$$\frac{1}{R}\Sigma_{r=1}^{R}\frac{1}{J}\Sigma_{j=1}^{J}\frac{1}{M}\Sigma_{m=1}^{M}[P(\theta_m|a_{jT},b_{jT},c_{jT})-P(\theta_m|\hat{a}_j^r,\hat{b}_j^r,\hat{c}_j^r)]^2h(\theta_m),$$

where $\hat{a}_j^r$, $\hat{b}_j^r$, and $\hat{c}_j^r$ are the item parameter estimates for the item j obtained from the replication r; a_{jT}, b_{jT}, and c_{jT} are the true item parameters for the item j; θ_m is the m-th proficiency when the interval (-4, 4) on the θ-scale is divided into M points; $h(\theta_m)$ is the weight proportional to the density function of N(0, 1); R is the number of replications ; and J is the number of unique (non-common)items on Form B. For this study, M and R were 31 and 100, repectively.

Results

Most of PARSCALE runs used in the FI method converged. However, for the 2PL model, the first PARSCALE run did not converg at six of the 100 replications in the 500 sample size for the N(1, 1) condition. For the 3PL model, it did not converge at one of the 100 replications in the 500 sample size for the N(1, 1) condition. For the 2PL model, the second PARSCALE run did not converge at one of the 100 replications in the 500 sample size for the N(1, 1) condition. For the 3PL model, it did not converge at one of the 100 replications in the 500 sample size for the N(0, 1) condition. The results for each replication in the second PARSCALE run were not used to calculate the criterion.

The values of the ICC criterion under 2PL model and 3PL model are presented in Figures 1 and 2, respectively. In each figure, the plots in the two rows give the results for the sample size of 3000 and 500.

The figures showed that the MSE decreased as the sample size increased from 500 to 3000. The MSE increased as the number of common items decreased from 20 to 4 and as the new group that took Form B became nonequivalent to the old group that took Form A with the distributions((N(0,

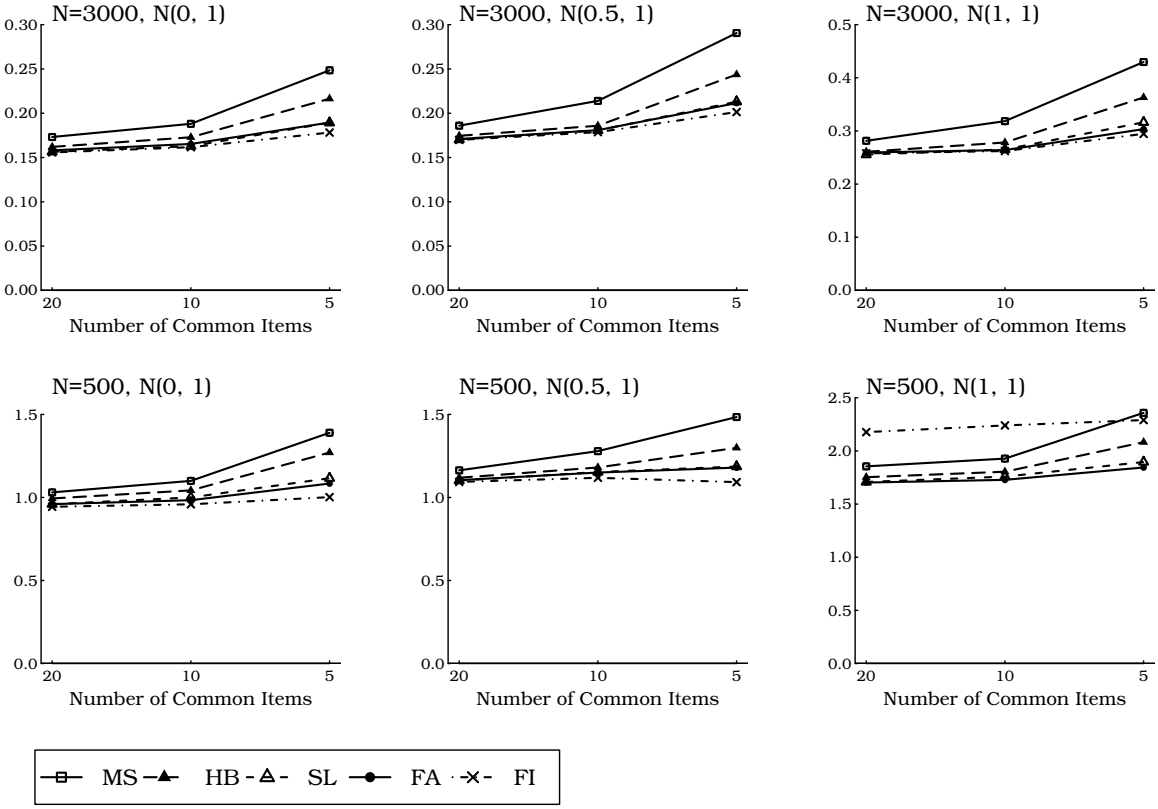

Figure 1: Averaged Mean Squared Error for the ICC criterion under 2PL model

1), N(0.5, 1), and N(1, 1)), with a few exceptions.

The MS method had the largest MSE, and the FI method had the lowest MSE, with a few exceptions. Among ICC calibration methods, the FA and SL methods showed similar performances and had lower MSE than the HB method under the 2PL model; whereas, the SL methods showed performances similar to the HB method, and the FA method had lower MSE than both the methods in the 3PL model.

Summary and Discussion

This study focused onvfive calibration methods (FA, SL, HB, MS, and FI) and used simulated data to compare their performances under various conditions.

Several results were found in this study. First, the MSE decreased as the sample size increased from 500 to 3000. Second, with a few exceptions, the MSE increased as the number of common items decreased (20, 10, and 4). Third, the MSE increased as the new group that took Form B became nonequivalent to the old group that took Form A(N(0, 1), N(0.5, 1), and N(1, 1)). These results agree with the previous studies.

The performances of five calibration methods were similar in each condition except for the FI method in the 500 sample size for the N(1, 1) condition under the 2PL model. In the 20 common items condition, there were small differences among the five calibration methods. As the number of common items decreased, the differences increased slightly. The MS method had the largest MSE in every condition. Under the 2PL model, the FI method had the lowest MSE except the 500 sample size with the N(1, 1) condition. Under the 3PL model, the FA and FI methods had the lowest MSE. The FI method and the FA method both performed well. However, compared to the FI method, the FA

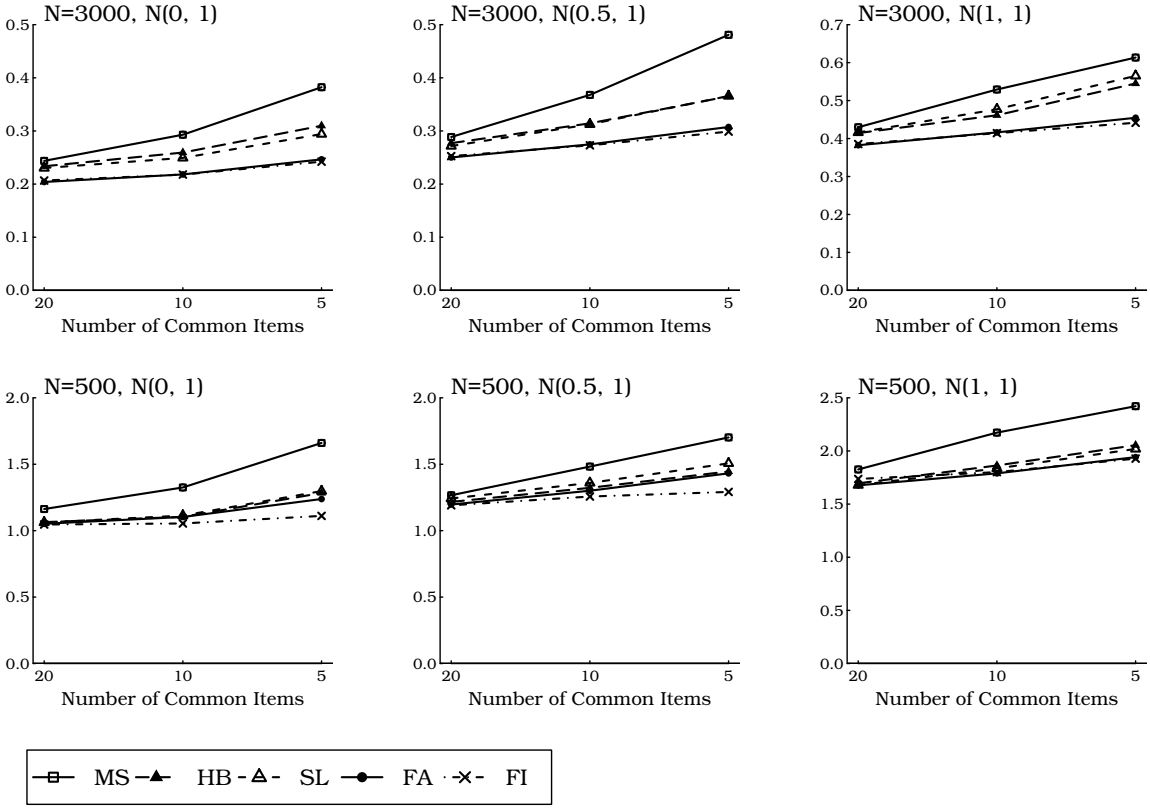

Figure 2: Averaged Mean Squared Error for the ICC criterion under 3PL model

method did not give any problems with the convergence of the computer program. From the practical viewpoint, the FA method might be more useful than the FI method.

In this study, the performances of five calibration methods were examined using simulated data. There are various other types of tests and examinations in practice. Our future research will cover more practical situations.

REFERENCES (RÉFERENCES)

Arai, S. & Mayekawa, S. (2011). A comparison of equating methods and linking designs for developing an item pool under item response theory. *Behaviormetrika*, *38*, 1-16.

Fujita, T., & Mayekawa, S. (to appear). Selecting the most appropriate IRT logistic model and method of equating for in-house listening pre-and post-tests. *Japanese Journal for Research on Testing*. (in Japanese)

Haebara, T. (1980). Equating logistic ability scales by a weighted least squares method. *Japanese Psychological Research*, *22*, 144-149.

Han, K. T. & Hambleton, R. K. (2010). WinGen 3 [computer program].

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*, 3-24.

Kim, S., & Lee, W.-C. (2006). An extension of four IRT linking methods for mixed-format tests. *Journal of Educational Measurement*, *43*, 53-76.

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices (2nd Ed.)*. New York: Springer-Verlag.

Lee, W.-C., & Ban, J.-C. (2010). A comparison of IRT linking procedures. *Applied Measurement in Education*, *23*, 23-48.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum.

Marco, G. L. (1977). Item characteristic curve solutions to three intractable testing problems. *Journal of Educational Measurement*, *14*, 139-160.

Mayekawa, S. (1991). Parameter Estimation. In Shiba, S. (Eds.), *Koumoku hannou riron –kiso to ouyou (Item response theory: bases and applications)* (pp. 87-129). Tokyo: The University of Tokyo Press (in Japanese).

Muraki, E., & Bock, R. D. (2003). PARSCALE 4 [computer program]. Chicago, IL: Scientific Software.

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

Zimowski, M. F., Muraki, E., Mislevy, R. J., & Bock, R. D. (2003). BILOG-MG 3 [computer program]. Chicago, IL: Scientific Software.