Supervisor Taehoon Kang

Thesis for the Degree of Master

A Comparison of MCMC and MMLE/EM Algorithms for Fixed Item Parameter Calibration

2013

The Graduate School of

Sungshin Women's University

Department of Education

Sohee Kim

A Comparison of MCMC and MMLE/EM Algorithms for Fixed Item Parameter Calibration

Supervisor Taehoon Kang

A Master Thesis submitted to
the Department of Education and the Graduate
School of Sungshin Women's Univesity

August 2013

The Graduate School of
Sungshin Women's University
Department of Education

Sohee Kim

Approval Sheet

This certifies that the masters thesis
of Sohee Kim is approved.

_____

Hyungjin Shim

_____

Taehoon Kang

_____

Myunglang Kim

The Graduate School of
Sungshin Women's University

KOREAN ABSTRACT

본 연구는 FIPC(Fixed Item Parameter Calibration) 방법으로 두 개의 검사간 문항모수 연계를 실시하고자 할 때 MCMC(Markov Chain Monte Calro) 알고리듬과 MMLE/EM(Marginal Maximum Likelihood Estimation using the Expectation-maximization) 알고리듬의 수행 결과를 비교하고자 하였다. 이들의 수행을 비교해봄으로써, MMLE/EM 알고리듬이 적용되기 어려운 상황에서 MCMC를 사용하여 IRT 동등화를 보다 쉽게 확장할 수 있을 것으로 기대한다.

본 연구에서는 다음과 같은 연구문제를 설정하였다.

1. 표본 크기에 따른 두 방법의 동등화 수행 결과를 비교한다.

2. 공통 문항 수에 따른 두 방법의 동등화 수행 결과를 비교한다.

3. 능력 분포에 따른 두 방법의 동등화 수행 결과를 비교한다.

4. 두 방법의 RMSE 값의 차이가 유의미한 지 확인한다.

본 연구를 위하여 각 조건에 맞는 문항 자료들을 생성하였다. 시뮬레이션 연구를 위한 시뮬레이션 조건으로 두 집단의 피험자 수(N=500, 2000)와 공통문항의 수(CI=10, 20, 40), 마지막으로 능력분포(N(0.0, 1.0), N(.25, 1.1^2),

N(.50, 1.2^2))를 고려하였다. 각 조건에서 10개의 자료를 반복적으로 생성하여 두 동등화 방법을 각각 적용하였고, 그 결과 문항 모수를 추정하였다. 각 조건에서 두 방법의 수행력을 비교하기 위하여 상관계수와 RMSE 값을 산출하여 비교하였고, 산출된 RMSE 값의 차이가 유의미 한지 판단하기 위하여 t 검증을 실시하였다.

본 연구를 통해 얻은 결과를 연구문제에 따라 정리하면 다음과 같다.

1. 표본크기가 500에서 2000으로 증가함에 따라서 MCMC와 MMLE/EM 알고리듬 모두 더 좋은 수행력을 보였으며, 두 방법 모두 높은 상관계수와 작은 RMSE 값을 산출하였다.

2. 공통문항의 수가 10, 20, 40으로 변화함에 따라서 MCMC와 MMLE/EM 알고리듬 모두 더 좋은 수행력을 보였으며, 두 방법의 결과 또한 유사했다.

3. 주어진 세 가지 능력분포에서 두 방법이 유사한 결과를 산출했으며, 능력분포에 따른 차이는 미미했다.

4. 각 조건별로 RMSE 값의 차이가 유의미한지 검증하기 위한 t 검증의 결과, 변별도 모수에서 유의미한 차이가 발견되었으며, 공통문항의 수가 40개 일 때 곤란도, 변별도 모수 모두에서 유의미한 차이가 발견되었다.

위 결과를 통하여, 두 개의 알고리듬이 문항 모수 연계를 위한 유용한 도구로 사용될 수 있음을 확인하였다. 또한 다양한 모형, 시뮬레이션 조건, 사전분포 등을 고려한 후속 연구가 수행될 필요가 있다.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURE

Ⅰ. Introduction

1.1 Research background

In the practice of educational situations, fair scores are assigned to an examinee through a test, and it is most important not to suffer from relative disadvantages when it comes to the interpretation of the test results. Moreover, if the results of the test are used for the purpose of selecting and evaluating an examinee, the objective interpretation of the results will be more important. Therefore, the procedures of equating, as one of the statistical manipulation forms, are required to avoid these relative disadvantages.

A test of equating involves a researcher estimating the ability and item parameters via alternative forms or parallel forms, which are used to measure the same constructs, and then exploring whether the ability and item parameters estimated by the previous test form correspond with certain values estimated by other test forms. In other words, equating is a statistical process that is used to adjust scores on test forms so that scores on the forms can be used interchangeably(American Educational Research Association, American Psychological Association & Nation Council on Measurement in Education, 1999; Kolen & Brennan, 1995). In these kinds of test equating settings, interpreting the results of the test with impartially and giving fair test scores to subjects have

become the most critical issues when the test is employed for the important purpose of admission into a university, such as the ACT(American College Testing) or SAT(Scholastic Aptitude Test) tests.

If there is no test equating, some examinees may take an easier test and get better scores, while other examinees may take a more difficult test and be relatively disadvantaged. If we can develop appropriate alternative forms or parallel forms all the time and evaluate students by tests that have the same characteristics, every examinee will be able to obtain unbiased outcomes. However, it is almost impossible to make a perfectly equivalent alternative test or parallel test, and as a result, a statistical manipulation called equating is essential.

In other words, when a standardized test is used to assess the growth of individual students or the effectiveness of educational organizations over time, it is necessary to put the scale of the item and ability parameter estimates onto the same metric across multiple test forms or years. Under item response theory(IRT), there are several linking procedures commonly used for this purpose such as separate calibration with linking, concurrent calibration, and fixed item parameter calibration(FIPC).

Previous studies related to FIPC had problems in the process of estimating parameters, and thus it was reported that the estimated parameters which were used in studies were less accurate and the FIPC method was criticized. However, researchers fixed the problems in implementing the FIPC method and could estimate ability and item

parameters properly. There is a significant study suggesting that FIPC tends to produce partial estimates(Baldwin, Baldwin, & Nering, 2007; Keller, Keller, & Baldwin, 2007; Paek & Young, 2005; Skorupski, Jodoin, Keller, & Swaminathan, 2003) compared to other linking procedures.

Furthermore, Paek and Young(2005) indicated that the use of a fixed prior ability distribution could cause a potential problem in FIPC through the marginal maximum likelihood estimation using the expectation-maximization(MMLE/EM) algorithm, and suggested a way to solve the problem. To address this potential problem, Kim(2006) mentioned statistical explanations for five fixed parameter calibration(FPC) methods, rooted in marginal maximum likelihood estimation through the EM algorithm and evaluated them. The five FPC methods described are distinguished from each other by how many times they update the prior ability distribution and by how many EM cycles they use(Kim, 2006). That is, he compared several possible FIPC procedures and identified one of them as a sound method that can provide accurate linking results. Recently, Kang and Petersen(2009) compared the correct FIPC procedure performed using the software PARSCALE with various linking methods such as concurrent calibration and separate calibration with linking, and showed that the FIPC method could be as good of an item-scale linking tool as the others. He suggested that the appropriate avenue should be used to yield more precise outcomes.

Under the Markov Chain Monte Carlo(MCMC), the idea associated

with updating the ability prior distribution through an iterative estimation procedure may be realized by using hyper-parameters for ability prior.

Considering that MCMC in IRT is increasingly used, it will be useful to check the utility of the MCMC algorithm for FIPC because the application can be easily extended to much more complicated situations where the MMLE/EM algorithm is very difficult to apply.

1.2 Research objectives

In the literature review above, no further studies investigated if there are differences between the MCMC and MMLE/EM algorithms in the procedure of equating, that is, if given data could be explained by both methods. Thus, the main purpose of this study is to show the differences in both methods with data collected by the NEAT data collection design through a simulation study. In other words, the main goal is to compare the performance of MCMC and MMLE/EM algorithms when linking between two test forms conducted with the FIPC method. Considering the increasing use of MCMC in IRT, the implications of whether or not MCMC is appropriate to apply with the FIPC method are important.

The specific objectives of this study comparing the MCMC and MMLE/EM algorithms were as follows:

1. To compare the results of the two methods as sample size changes

2. To compare the results of the two methods as the number of common or fixed items changes

3. To compare the results of the two methods as the true target distribution changes

4. To investigate whether the difference in RMSE values produced by the two methods is significant

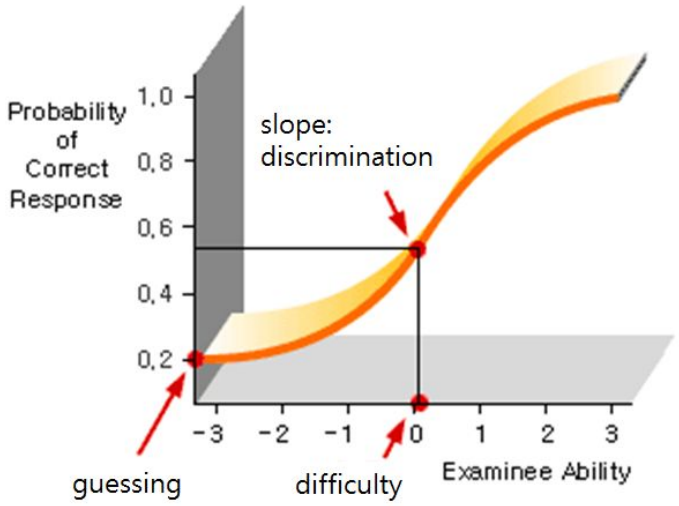
To answer these questions, a simulation study was conducted. The generating item parameters were selected from calibrations of a real data set. We used two forms that consisted of a 50-item multiple-choice test with a set of items in common. Also, the data was collected through a non-equivalent group anchor-test data collection design. And then, the procedures of equating were conducted with the FIPC method in each condition. Finally, to answer research question 4, RMSE values derived from both the MMLE/EM and MCMC algorithm under the same conditions were compared.

Ⅱ. Theoretical Background

2.1 Item Response Theory

2.1.1 Definition of Item Response Theory

Item Response Theory is analyzed by the item characteristic curve presenting the unique features of each item, unlike classical test theory, which analyzes tests or items by the total score of the test(Seong, 2001). The item characteristic curve(ICC) including item difficulty, item discrimination, and item guessing is as follows.



[Figure 1] Item Characteristic Curve(ICC)

There are two assumptions of IRT: one is the unidimensional assumption and the other is the local independence assumption. First, the unidimensional assumption means that one test measures only one construct. One test measures a latent variable, a trait of a human being and the latent variable measured by a test should be related to only one characteristic.

Next, the local independence assumption means that the response of one item by certain subjects with some extent of ability does not affect the response of other items(Han, 2011). That is, the probability of choosing the correct option in item 1 does not affect the probability of answering correctly in item 2.

In IRT having these assumptions, an item parameter is defined by a population, not a sample for a certain test, and the ability parameter of the examinee is defined by an item pool, not part of the item pool collected for a particular test. Thus, the item-parameter and person-parameter are independent of each other, and both parameters are put on the same scale. Item Response Theory is categorized into a dichotomous IRT model and a polytomous IRT model with a criterion of the examinee's response category.

A dichotomous IRT model is used in this study, and thus only descriptions about a dichotomous IRT model were referred to. A dichotomous IRT model is used when the response of subjects is presented by two types. In a multiple choice item a common practice is to include two possible responses: "correct(1)" and "wrong(0)". These

dichotomous IRT models are classified into one-parameter logistic models, two-parameter logistic models and three-parameter logistic models, according to the number of item parameters used in the model.

To carry out the study, choosing the kind of model is a crucial issue, and researchers should select a model that is the most suitable for the given data. A dichotomous IRT model is largely categorized into a normal ogive model and a logistic model with respect to functions used in the model, and the logistic model is more preferred than the normal ogive model, which has intricate procedures. Thus, the two-parameter logistics model is used in this study, too. However, the IRT that has a unidimensional assumption is difficult to apply to various situations because of the strong assumption, and this is required of a large sample. Also, the description of cognitive process is insufficient in IRT. Therefore, IRT is expanded into multidimensional IRT, mixture IRT, multilevel IRT and cognitive diagnostic models to make up for the disadvantages I mentioned.

2.1.2 One-Parameter Logistic Model(1PLM)

The one-parameter logistic model(1PLM) includes only item difficulty. In other words, 1PLM implies that item discrimination is equal across all items, and all item guessing parameters are zero. In the item characteristic curve under IRT, when the model does not include item discrimination and guessing, item difficulty means ability level(θ), the

points at which the curves attain a 50% probability of their respective items being passed. As a result, the higher values of item difficulty the more difficult the item. It is expressed as follows:

$$P(\theta_j) = \frac{1}{1 + e^{-1.7(\theta - b)}}$$

θ = ability level of examinee

b = item difficulty parameter

$P(\theta_j)$ = probability that subject with ability Θ responds to item j

2.1.3 Two-Parameter Logistic Model(2PLM)

Unlike the one-parameter logistic model, the 2PLM includes only the item difficulty parameter and it is assumed that all discrimination parameters are equal. The two-parameter logistic model lies under the assumption that every item has a different discrimination parameter. Item discrimination means a slope at the spot indicating item difficulty and the higher discrimination parameter means the item is better to distinguish subjects with high ability and low ability. Item discrimination is literally the extent to which it distinguishes among examinees having ability differences, and the item should be eliminated in the test if the discrimination parameter has a negative number. The formula of the two-parameter logistic model including item difficulty and discrimination

is as follows:

$$P(\theta_j) = \frac{1}{1 + e^{-1.7a(\theta - b)}}$$

θ = ability level of examinee

b = item difficulty parameter

a = item discrimination parameter

$P(\theta_j)$ = probability that subject with ability Θ responds to item j

2.2 Equating

2.2.1 Definition of Equating

As mentioned above, test equating is aimed at finding out the corresponding score that means the same level of ability between test X and Y to measure the same construct. That is, test equating is to adjust the difference in the level of difficulty between the two tests statistically. Test equating is required for scale tests, such as TOEFL(Test of English as a Foreign Language), SAT(Scholastic Aptitude Test), ACT(American College Test), NAEP(National Assessment of Educational Progress) and NAEA(National Assessment of Educational Achievement)and is used in multiple test forms. Also, in educational

situations, horizontal equating, linking different groups that involve subjects with similar or the same ability, or vertical equating, examining educational growth, are required. In vertical equating, more complicated procedures are required since the levels of examinees' ability as well as test difficulty are different(Nam, 2001). The necessary conditions for these test equatings are as follows.

First, equity is needed for test equating. Equity means that every student can get impartial scores regardless of the kinds of tests taken by examinees. Therefore, two tests that measure different constructs or characteristics cannot be equated and the tests with different reliabilities also cannot perform equating. Moreover, those that have different reliability cannot be equated. Also, only perfectly reliable alternative tests or parallel tests can meet the condition of equity and link with each other. Ironically the two tests are met perfectly the condition of equity, the two tests do not need to equate. As I mentioned above, because it is impossible to produce a complete alternative test or parallel test, eventually, test equating is concluded by the extent of equity.

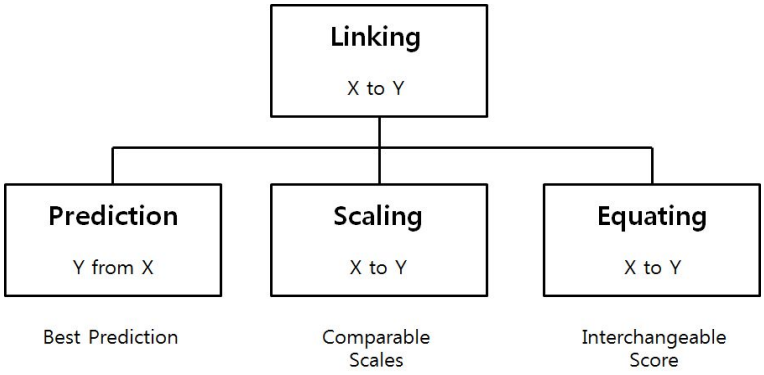
Second, symmetry is needed for test equating. Symmetry means that the opposite situation should be satisfied when a score, x, of test X corresponds to a score, y, of test Y. Therefore, for example, regression analysis cannot be an appropriate method for equating, because it is against symmetry equity.

Thus, the method to link scores obtained by two tests is to divide them into three parts including equating as shown in Figure 2

(Pommerich, 2011). First, predicting estimates of an examinee's score based on other information about the examinee in one test. For example, one's scores on a science test are predicted from scores on a math test. However, this method cannot meet the condition of comparability because corresponding scores are different between when test X is predicted by test Y and when test Y is predicted by test X. Thus, it is impossible to interpret linked scores by this method as comparable or interchangeable scores.

The second linking method is scaling, and this method considers that two of the scores are comparable, if both tests' scores are placed on the same scale. These scaling methods can be divided into several types according to the criteria including homogeneity or heterogeneity of the constructs.

Finally, equating means a statistical moderation in order to interpret the scores as comparable and interchangeable scores, when the alternative test was conducted to measure the same construct. According to Lord(1980), these procedures of equating are possible when two tests are alternative tests or parallel tests which are identical in content and metrology.



[Figure 2] Linking

2.2.2 Type of Equating

Test equating can be classified into a linear or equipercentile approach based on classical test theory or item response theory. Traditionally, linear and non- linear methods have been considered under the CTT. One is the linear approach and the other is the equipercentile approach(Budescu, 1985; Angoff, 1971; Braun & Holland, 1982; Kolen & Brennan, 2004; von Davier, 2008). These two methods assume that the ability distributions of two examinee groups having taken each test are equivalent.

In other words, both equating methods based on the CTT have a hypothesis that each group is randomly equivalent, under the hypothesis; the differences in the test score distributions are stems from the differences in the difficulty of the tests. Therefore, traditional equating

methods can be used, only when data is collected by a random group design or a single group design.

First, in linear equating, scores from different tests are regarded to be identical when the standard scores(z score) on one test are the same as that on another test. Stated differently, the scores that have the same mean and standard deviation will correspond with each other. If score distributions of both tests are similar, accurate equating results are yielded even if the sample sizes in the tests are small.

Next, equipercentile considers scores that are in the same percentile in each test as the corresponding scores. Thus, the scores having the same percentile in each test's score distributions correspond to each other. The equipercentile method can get accurate results, even if the two test score distributions are a little bit different when the sample size is large. However, smoothing is required to smooth the cumulative distribution curve when the sample size is small.

Unlike the traditional equating method mentioned above, the equating method based on IRT can apply to situations where the equating method based on CTT is hard to apply. When test equating based on IRT is conducted, the situation where we know neither item parameters nor ability parameters is common. Thus, the problem of scale indeterminacy occurs, and in most commercial programs, mean and standard deviation of the ability distribution are fixed to 0 and 1, respectively. Like this, when the person-parameters of both groups are not known, item parameters measured independently from each data have invariability

only until a linear transformation, and the linear transformation, which replaces a scale obtained from one set of data with another scale obtained from another set of data is required in order to put estimated item parameters on the same scale and compare them. That is, item-parameter and person-parameter calculated from test X and Y, respectively, have the following relationship, even if they stem from the same item.

$$b_Y = Ab_X + B$$
$$a_Y = a_X / A$$
$$c_Y = c_X$$
$$\theta_Y = A\theta_X + B$$

Thus, we can calculate linking coefficients(slope A and intercept B that represent a linear transformation) to conduct a linear transformation. In the NEAT design used in this study, many methods, such as the Mean/Sigma method, Mean/Mean method, Stocking Lord method and Haebara method can be employed to carry out separate calibrations with linking.

First, linking coefficients are calculated by using the mean and standard deviation of common item difficulty in the Mean/Sigma method.

When difficulty parameters of common items in the two tests are called b_{Xs} and b_{Ys}, respectively, the process for calculating the linking coefficients is expressed as follows:

$$A = \frac{S_Y}{S_X}\ ,\ B = \overline{b_Y} - A\overline{b_X}$$

Second, linking coefficients are computed using the mean of the common item difficulty as well as discrimination in the Mean/Mean method. And the process for calculating linking coefficients is expressed as follows:

$$A = \frac{\overline{a_X}}{\overline{a_Y}}\ ,\ B = \overline{b_Y} - A\overline{b_X}$$

Third, the Stocking Lord method called the Test Characteristic Curve method(TCC method) calculates linking coefficients that minimize the sum of squares of differences between two curves, after drawing up test characteristic curves(τ_X , τ_Y) with common items in each test. When the TCC of a new test X is made, the following function, F, is presented by a relationship between A and B, because the item-parameter and person-parameter of common items reflect linking coefficients.

$$F = \sum_{j=1}^{N} (\tau_{Xj} - \tau_{Yj})^2$$

Finally, the Haebara method is called an Item Characteristic Curve method(ICC method). Unlike the Stocking Lord method using the test

characteristic curve, this method uses the item characteristics curve, and the function is as follows:

$$H = \sum_{i=1}^{CI} \sum_{j=1}^{N} (p_{\Xi j} - p_{Yij})^2$$

CI = the number of common item

p_{Xij} = ICC with item I, subject j takes in test X

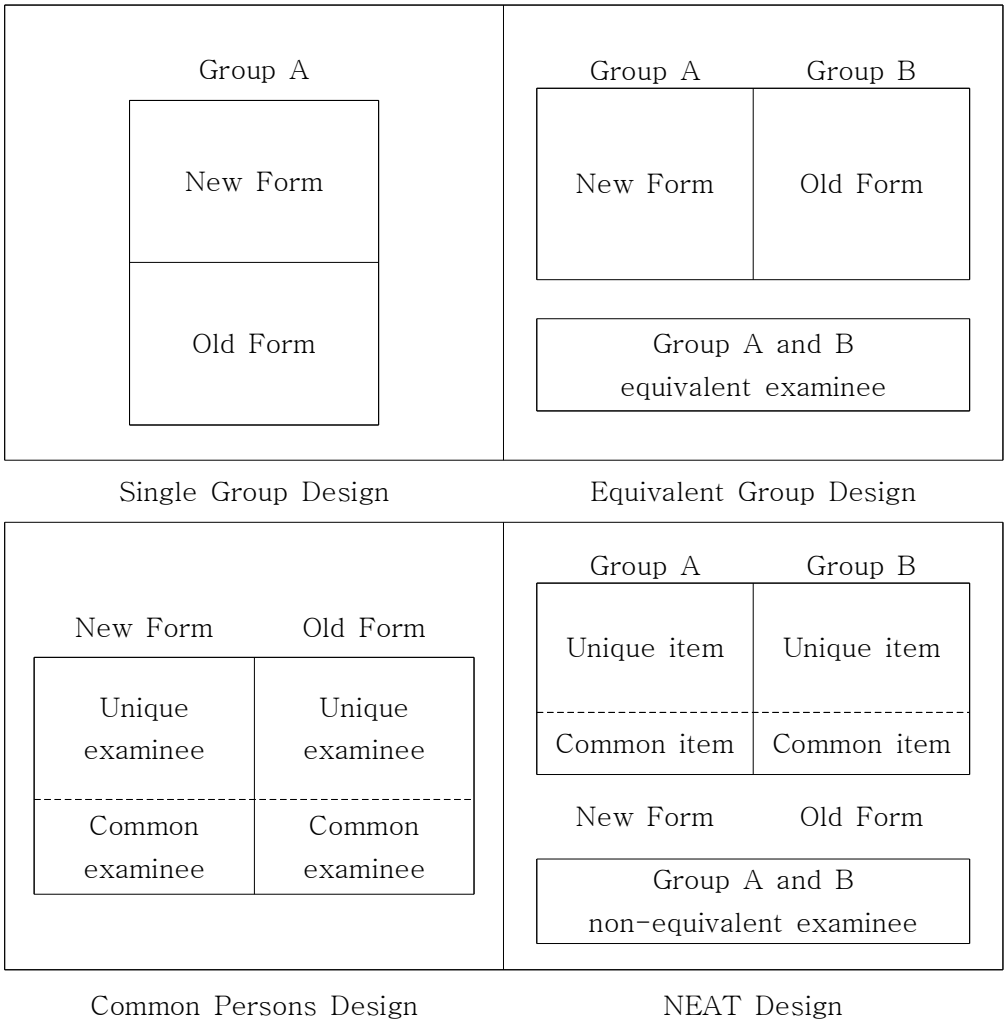
There are more equating methods based on IRT than equating methods using liking coefficients. There are ways to equate two tests without linking coefficients; namely concurrent calibration and Fixed Item Parameter Calibration(FIPC), which is used in this study. Especially, FIPC method conducted in this research is to fix the common item parameters and let newly estimated item parameters put on common item scale. Compared to methods using linking coefficients or concurrent calibration, FIPC is useful for managing item parameters. Also, the accuracy of FIPC results has been proved in previous studies.

2.2.3 Designs of Equating

A set of equating designs is required to collect data for equating. In order to choose a proper design method, there are several things to consider, such as the characteristics of the examinee group, the number

of tests, and usage of the anchor test. In general, common subjects or items included commonly or two groups with equal ability are required for data collecting design, which is appropriate for test equating. Data collection design is a practical issue and also a statistical issue, in the sense that we should choose an equating method according to a data collection design.

Data collection design is categorized as follows: (1) single group design conducted on the same examinee group with two tests, (2) Equivalent group design performed on two groups with equal ability, (3) Common persons design conducted on two groups with non-equivalent ability, but sharing some portion of subjects who respond to both tests and (4) NEAT design(Anchor test design) conducted on two groups with non-equivalent ability, but sharing common items. Figure 3 was revised from Kolen and Brennan(2004).



[Figure 3] Equating Design

First, single group design is the simplest data collection design. In this data collection design, it satisfies the assumption of ability equivalency, since both the old test and new test are taken by the same examinees. When using this design, the difference between group-level

performances on the two forms is taken as a direct indication of the difference in difficulty between the forms(Kolen & Brennan, 2004).

Second, the equivalent group design is to administer a test in two groups with the same ability. Unlike single group design, examinees do not need to take two test forms in the equivalent group design. Also, the equivalent group design requires the same conditions of time and sample size. Such requirements might be difficult to be met in some situations and exacerbate concerns about test form security(Kolen & Brennan, 2004). Thirdly, common persons design is a way to share some of the examinees who are taking two different tests when both tests were implemented by two groups, respectively. The advantages of a common person design are that a smaller sample of students is needed and it is very powerful at detecting differences. The disadvantages of a common person design are that examinees are required to test twice, and factors associated with testing twice, such as motivation, may influence test performance(Wan et al, 2009).

Finally, a mixture of group difference and test form difference exists in the difference in group-level administration, because NEAT design does not assume equivalent groups. Therefore, the effort is needed to separate group difference from test form difference. As a result, fixed or common items are required in the test. The NEAT design is the most common design in equating, because one subject need take only one test form and a large sample is not required in NEAT design. Moreover, if an external anchor test is used that does not contribute to the

examinee's score, unique items that contribute to the score can be disclosed following the test data(Kolen & Brennan, 2004).

2.3 MCMC and MMLE/EM Algorithms

2.3.1 Markov Chain Monte Carlo(MCMC) Algorithm

A recent survey places the Metropolis algorithm among the 10 algorithms that have had the greatest influence on the development and practice of science and engineering in the 20th century(Beichl&Sullivan, 2000). This algorithm is an instance of a large class of sampling algorithms, known as Markov Chain Monte Carlo(MCMC). These algorithms have played a significant role in statistics, econometrics, physics and computing science over the last two decades. There are several high-dimensional problems, such as computing the volume of a convex body in d dimensions, for which MCMC simulation is the only known general approach for providing a solution within a reasonable time(polynomial in d) (Dyer, Frieze, & Kannan, 1991; Jerrum & Sinclair, 1996).

The Markov Chain, developed by mathematicians in Russia, is one of the methods of decision making and a useful tool to study change or developments of a certain system. It is an analysis method based on the stochastic process that any events or test results were determined by

immediate prior events or experimental results. Also, the Markov Chain analyses the change or progress of the specific system by using the probabilities (or transition matrix) change from one situation to a different situation. As a result, the Markov Chain is a stochastic technique to predict changes from an original state to the next state. Only the events that occur in the previous decide the next events and the value of results should be one of the discrete random variables.

Monte Carlo is a kind of simulation technique used to obtain the probability distribution of a value that you want to know from statistics in a repeatable experiment. In other words, it is an analysis that relies on random variables, and specific values or probability distributions are obtained by inversing the statistical data based on a plethora of experiments. Due to the nature of Monte Carlo, the more statistical data and the more regular the input values you have, the more precise the results you can get.

Another feature of Monte Carlo is that it is easy to apply. So, the value of results can be obtained by inputting short codes of computer programs, without various or complex procedures. This advantage is a really big help when you need to calculate values that are difficult to get theoretically. In recent years, with the development of analysis using computers, Monte Carlo has been used in the fields of science and engineering.

As a result, Markov Chain Monte Carlo is a method to find a proper fair sampling reflected in the probability distribution function, when

given the specific probability distribution function. Of course, different ways can be used; however, Markov Chain Monte Carlo is a common analysis method when there are a lot of variables or the probability distribution function is not a normal distribution. WinBUGS is statistical software for Bayesian analysis using Markov Chain Monte Carlo (MCMC) methods.

2.3.2 MMLE/EM Algorithm

Bock and Lieberman(1970) proposed marginal maximum likelihood estimation(MMLE) to remove the effect of incidental parameters by integrating them over the θ distribution. MMLE estimates item and θ parameters in separate stages without a back and forth scheme; although this approach poses a formidable computational task and is practical for only very short tests(Seong, 1990).

In order to solve the computational problem, Bock and Aitkin(1981) used a modification of the EM algorithm formulated by Dempster, Laird, and Rubin(1977). One characteristic of this MMLE/EM approach is the use of Bayesian concepts, such as prior and posterior distributions(Seong, 1990). Basically, the MMLE method is different from the Bayesian method, but the term "marginal" is related to integration and includes Bayes' theorem naturally. Bayes' theorem is a procedure used to improve the initial probability of certain cases by using data or

information from experiments. The equation is as follows:

$$P(B|A) = \frac{P(A|B)P(B)}{\int P(A|B)P(B)dB} = \frac{P(A|B)P(B)}{P(A)}$$

The MMLE method implemented this Bayes theorem expressed as follows:

$$P(\theta_j|u_j,\tau,\xi) = \frac{P(u_j|\theta_j,\xi)P(\theta_j|\tau)}{\int P(u_j|\theta_j,\xi)P(\theta_j|\tau)d\theta_j} = \frac{P(u_j|\theta_j,\xi)P(\theta_j|\tau)}{P(u_j|\tau,\xi)}$$

The u_j is the item response pattern of examinee j, the τ is population parameter and the ξ is the item parameter in the above formula. In the expression mentioned above, the distribution defined by the population parameter is a prior distribution and one subject's marginal probability of an item response pattern can be calculated in accordance with the information of the prior distribution and item parameter. Thus, each response pattern has a posterior distribution and quadrature points are employed for convenience of integration.

Also, to understand the MMLE of Bock and Aitkin(1981), one should be aware of the notion about X_k, the number of frequency examinees in a population of size N expected to have ability score X_k, and $\overline{f_{ik}}$, the number of examinees in the population at ability X_k expected to respond correctly to the item i(Baker & Kim, 2004). The equations can be

written as follows:

$$\overline{f_{ik}} = \sum_{j=1}^{N} P(X_k|u_j, \tau, \xi)$$

$$\overline{r_{ik}} = \sum_{j=1}^{N} u_{ij} P(X_k|u_j, \tau, \xi)$$

An additional algorithm is required because $\overline{f_{ik}}$ and $\overline{r_{ik}}$ depend only on item parameters in MMLE. Then the EM algorithm is added to the MMLE and the procedure(Baker & Kim, 2004) is as follows:

1. Calculate $\overline{f_{ik}}$ and $\overline{r_{ik}}$ by using the initial value of the item parameters(E-step).
2. Estimate the item parameter through Newton-Raphson(NR) based on $\overline{f_{ik}}$ and $\overline{r_{ik}}$(M-step).
3. Compute a new value of $\overline{f_{ik}}$ and $\overline{r_{ik}}$ by implementing the assembled parameter estimates of each item(E-step).
4. As process 3 mentioned above, produce new item parameters(M-step).
5. If the overall likelihood is unchanged from the previous EM cycle, the item estimation process has converged and the process is terminated. Otherwise, E-step and M-step are repeated(Baker & Kim, 2004).

The MMLE/EM approach developed by the process mentioned above of Bock and Aitkin has been implemented in the BILOG computer program (Mislevy & Bock,1982, 1984, 1986). Because of the difficulty of direct integration with a digital computer, MMLE employs numerical analysis techniques to integrate examinees' θs over the θ distribution. BILOG has an option that allows the user to specify the number of quadrature points, values of quadrature points, and weights corresponding to each quadrature point.

Several item-recovery studies(Drasgow, 1989; Mislevy & Stocking, 1989; Qualls & Ansley, 1985; Yen, 1987) have investigated the accuracy of MMLE/EM for item and θ parameter estimation. Basically, accuracy was defined by how far the item and θ estimates were from their underlying parameters, and how these estimates were correlated with these parameters. Qualls and Ansley(1985), Yen(1987), and Mislevy and Stocking(1989) compared LOGIST and BILOG results. Qualls and Ansley found that BILOG uniformly took more time for calculating to produce estimates than did LOGIST, but the BILOG estimates were uniformly more accurate than the LOGIST estimates. Also, a number of studies were conducted to compare between LOGIST and BILOG.

The several recovery studies concluded that the BILOG estimates generally were more accurate than the LOGIST estimates for short tests and small numbers of examinees. However, none of the above mentioned studies specified different prior θ distributions or other factors associated with these distributions; only normal prior θ distributions were used.

The characteristics of the prior θ distribution affect the posterior distribution of θ used in estimating item parameters, and if the prior θ distribution is correctly matched with the underlying θ distribution, MMLE will produce consistent estimates of item parameters(Harwell, Baker, & Zwarts, 1988). Thus, the agreement of the specified prior and underlying θ distributions needs to be considered; the goal of this research was to investigate the role of the prior θ distribution in the parameter estimation under the MMLE/EM approach.

In addition, according to Baker and Kim(2004) there are several characteristics of the MMLE/EM algorithm.

1. The marginalization over θ produces estimates of item parameters that are consistent for tests of finite length(assuming that the IRT model and the ability population model are correct).
2. Conceptualizing subjects as a random sample form, a population with ability distributed in accordance with a density function $g(\theta|\tau)$ allows one to assume an arbitrary distribution of θ in the population sampled. Consequently, the metric of the item parameter estimates is defined by the location and scale parameters in τ.
3. It permits the imposition of a "Bayesian-like" structure on the estimation process such that inferences about θ are improved. As noted earlier, the EM algorithm illustrated here does not involve a classical Bayes' solution since the item parameters are treated as constants, meaning the parameter estimates are generated using

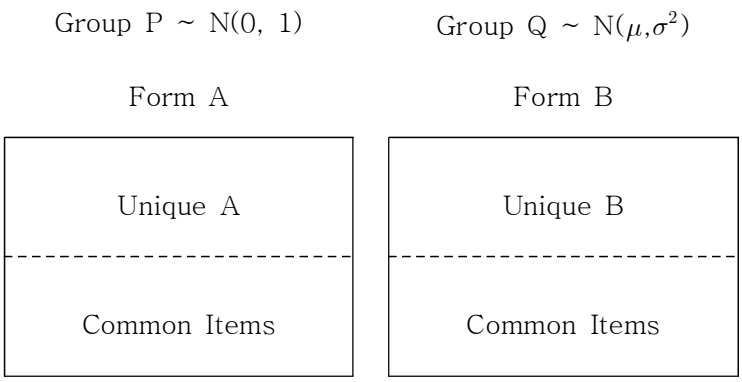
maximum likelihood methods.

4. The computer programming needed to implement an MMLE/EM approach is actually quite simple although the underlying mathematics are rather sophisticated. The E-step involves a very large number of computations even for a small number of items, but the computations are well structured and can be done in an orderly fashion.

Ⅲ. Method

3.1 Data and study design

In this study, the two calibration procedures are compared using simulated data. One is to use the FIPC through MMLE/EM algorithm performed by PARSCALE, and the other is the FIPC through MCMC algorithm performed by WinBUGS. However, the generating item parameters were selected from calibrations of real data sets(two 60-item mathematics achievement test forms with sample sizes ranging from 4,294 to 4,557). It is assumed that we have data for two forms of a 50-item multiple-choice test and that the two forms have a set of items in common(number of common items equals 10, 20 or 40). It is further assumed that the data were collected via a non-equivalent groups anchor-test(NEAT) data collection design(see Figure 1). Item parameters for the old form(Form A) already exist from the calibration of data from a previous administration. The task of interest here is to estimate item parameters of the new form and link them to the scale of the old form items.



[Figure 4] NEAT Data Collection Design for Common Item Equating

In the NEAT design, one test form is administered to one group of test-takers and another test form is administered to another group of test-takers. The two groups are naturally occurring or non-equivalent and therefore likely to differ in ability. For example, one might be a group taking the test in the fall and the other a group taking the test in the spring. A common test or anchor test, in this case common items, is administered to both groups in order to estimate the performance of the combined group on both forms, thus simulating, by statistical methods, the situation in which both groups take both forms.

The two parameter logistic model(2PLM) is used, and the distribution of the base group(Group P) is assumed to be N(0,1). Other factors included in the simulation design were (1) two different sample sizes for both the base and target groups(500 and 2000) and (2) three different ability distributions for the target group [N(0.0, 1.0), N(.25, 1.1^2), N(.50,

1.2^2)]. These factors are critical ones that affect calibration in practice.

The sample size of 2,000 was chosen to represent the usual practice of calibrating operational items with relatively large samples in order to produce stable item parameter estimates. The sample size of 500 was chosen to represent the minimum sample size in practice that is likely to yield acceptable calibration results. To generate old group(Group P) examinees, N(0, 1) distribution was used. The new group(Group Q) distributions were selected to represent those situations were the old and new groups are very similar in ability [N(0.0, 1.0)], the old and new groups differ somewhat in ability [N(.25, 1.1^2)], and the old and new groups differ significantly in ability [N(.50, 1.2^2)]. In practice, we seldom see group differences as extreme as those represented by the N(.50, 1.2^2) distribution.

<Table 1> 18 simulation conditions

N = 500		N = 2000	
# FI =10	N(0.0, 1.0) N(.25, 1.1^2) N(.50, 1.2^2)	# FI =10	N(0.0, 1.0) N(.25, 1.1^2) N(.50, 1.2^2)
# FI =20	N(0.0, 1.0) N(.25, 1.1^2) N(.50, 1.2^2)	# FI =20	N(0.0, 1.0) N(.25, 1.1^2) N(.50, 1.2^2)
# FI =40	N(0.0, 1.0) N(.25, 1.1^2) N(.50, 1.2^2)	# FI =40	N(0.0, 1.0) N(.25, 1.1^2) N(.50, 1.2^2)

Finally, there are a total of 18 conditions simulated in this study(2 sample sizes x 3 numbers of fixed items x 3 target group distributions) as shown in Table 1. Ten replications are generated for each condition for both base and target groups.

3.2 IRT calibration and Linking method

In this study, FIPC method was used for equating. First, item parameters were calculated in each group(group P and Q in Figure 1) respectively, unlike with separate calibration that needs a linear transformation for linking, there is no process such as linear transformation. Instead of the procedure of linear transformation, common or fixed item parameters in new group were replaced by item parameters were estimated in old group to estimate unique item parameters in new group. That is, parameter estimates in both groups are placed at the same scale by fixing common or fixed item parameters in new group.

As two studies conducted by Kim(2006) and Kang & Petersen(2009) mentioned above, this study tried to employ correct usage for linking through FIPC. Thus, 'FREE=(NOADJUST, NOADJUST)' and 'POSTERIOR' were used in the CALIB command to estimate more accurate item and person parameters via correct FIPC. 'FREE=(NOADJUST, NOADJUST)' is used to prevent rescaling of parameter and 'POSTERIOR' enables updating or the prior ability

distribution multiple times. So, careful usage about two commands is needed in PARSCALE.

3.3 Evaluation criteria

The accuracy or performance of the four IRT linking procedures will be evaluated with 1) the Pearson's product moment correlations and 2) root mean square errors between true and estimated and linked item parameters.

To evaluate the recovery of both methods, MCMC and MMLE/EM algorithms, the mean and standard deviation of correlations between true and estimated and linked item parameters were calculated in each condition. A better performance of FIPC was able to indicate a higher correlation between true and estimated item parameters and Pearson's product moment correlations is defined as follow:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x-\bar{x})(y-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x-\bar{x})^2 \sum_{i=1}^{n}(y-\bar{y})^2}}$$

Next, the degree of recovery in two methods was evaluated via root mean squared error(RMSE). RMSE is a intuitional and meaningful evaluation criteria and is used to deal with the differences between the

value estimated by model and the value observed. These RMSE value was calculated for two item parameter(a,b). For instance, RMSE is defined as follow when difficulty parameters were estimated.

$$SE(\hat{b_j}) = \sqrt{\frac{1}{n}\sum_{r=1}^{n}(\hat{b_{jr}} - b_j)^2}$$

Ⅳ. RESULTS

For each group P and Q, ten datasets were generated under each of 18 simulated conditions in this study. And the two different FIPC procedures were applied to them.

4.1 The Pearson's product moment correlations

Table 2 and Figure 3 involve the result for how strong the correlation between generating and estimated parameters were high by each of the two FIPC procedures. Especially the correlation between true and estimated difficulty parameters(b-parameters) were very high. In addition, both FIPC procedures showed positive and comparable correlation results and the they tend to get better as the sample size changes from 500 to 2000. For sample size N=500, the average means and SDs for the correlations between true and estimated discrimination parameters(a-parameters) across two FIPC procedures ranged between .845 and .934 and between .022 and .088, respectively. And in case of b-parameters, the correlation means and SDs ranged between .994 and .997 and between .001 and .006, respectively. For sample size N=2000, the average means and SDs for the correlations of a-parameters across two FIPC procedures ranged between .959 and .982 and between .005 and

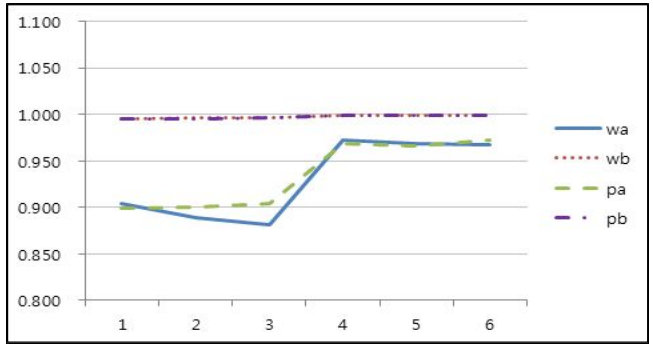
.023, respectively. And in case of b-parameters, the correlation means are observed that all values are .999 and SDs ranged between .000 and .001.

<Table 2> Average Means and Standard Deviations of the Correlation between Generating and Estimated Parameters of the Target Group (a-parameter)

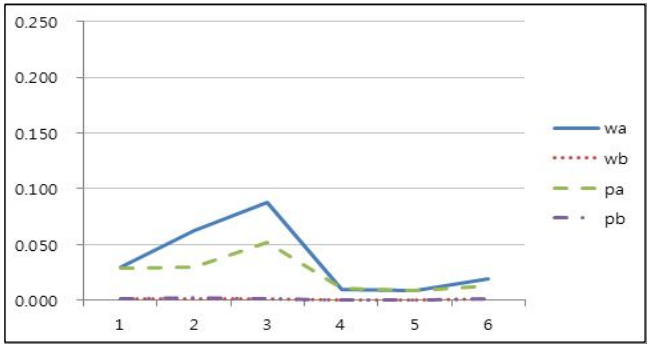
sample sizes	# common or fixed items	True Target Distributions	MCMC algorithm (WinBugs)		MMLE/EM algorithm (PARSCALE)	
			a-parameter			
			mean	SD	mean	SD
N=500	#FI=10	N(0.0,1.0)	0.904	0.030	0.899	0.029
		N(.25,1.1^2)	0.874	0.061	0.887	0.027
		N(.50,1.2^2)	0.845	0.041	0.912	0.044
	#FI=20	N(0.0,1.0)	0.889	0.063	0.900	0.030
		N(.25,1.1^2)	0.889	0.042	0.895	0.041
		N(.50,1.2^2)	0.894	0.030	0.900	0.034
	#FI=40	N(0.0,1.0)	0.881	0.088	0.904	0.052
		N(.25,1.1^2)	0.897	0.060	0.905	0.052
		N(.50,1.2^2)	0.921	0.032	0.934	0.022
N=2000	#FI=10	N(0.0,1.0)	0.972	0.010	0.969	0.011
		N(.25,1.1^2)	0.971	0.008	0.969	0.007
		N(.50,1.2^2)	0.977	0.006	0.976	0.005
	#FI=20	N(0.0,1.0)	0.968	0.009	0.966	0.009
		N(.25,1.1^2)	0.980	0.007	0.979	0.007
		N(.50,1.2^2)	0.972	0.008	0.972	0.007
	#FI=40	N(0.0,1.0)	0.967	0.019	0.973	0.013
		N(.25,1.1^2)	0.973	0.013	0.982	0.012
		N(.50,1.2^2)	0.959	0.023	0.975	0.016

<Table 3> Average Means and Standard Deviations of the Correlation between Generating and Estimated Parameters of the Target Group (b-parameter)

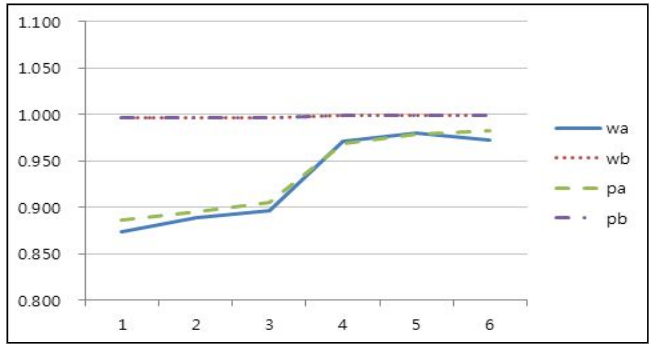
sample sizes	# common or fixed items	True Target Distributions	MCMC algorithm (WinBugs)		MMLE/EM algorithm (PARSCALE)	
			(PARSCALE)			
			mean	SD	mean	SD
N=500	#FI=10	N(0.0,1.0)	0.995	0.001	0.995	0.001
		N(.25,1.1^2)	0.996	0.001	0.996	0.001
		N(.50,1.2^2)	0.996	0.003	0.994	0.006
	#FI=20	N(0.0,1.0)	0.996	0.002	0.996	0.002
		N(.25,1.1^2)	0.996	0.001	0.996	0.001
		N(.50,1.2^2)	0.996	0.002	0.996	0.002
	#FI=40	N(0.0,1.0)	0.997	0.001	0.997	0.001
		N(.25,1.1^2)	0.996	0.002	0.996	0.002
		N(.50,1.2^2)	0.996	0.003	0.996	0.003
N=2000	#FI=10	N(0.0,1.0)	0.999	0.000	0.999	0.000
		N(.25,1.1^2)	0.999	0.000	0.999	0.000
		N(.50,1.2^2)	0.999	0.000	0.999	0.000
	#FI=20	N(0.0,1.0)	0.999	0.000	0.999	0.000
		N(.25,1.1^2)	0.999	0.000	0.999	0.000
		N(.50,1.2^2)	0.999	0.000	0.999	0.000
	#FI=40	N(0.0,1.0)	0.999	0.001	0.999	0.001
		N(.25,1.1^2)	0.999	0.001	0.999	0.001
		N(.50,1.2^2)	0.999	0.000	0.999	0.000



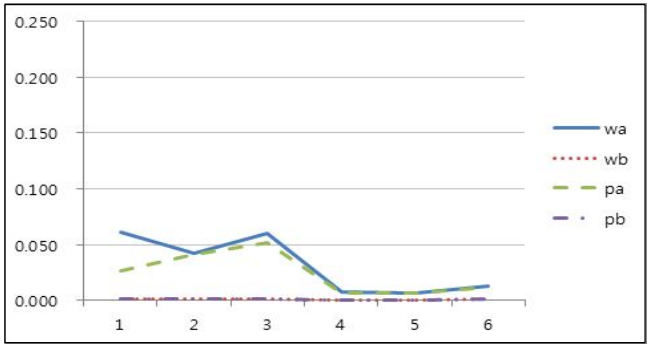
(a) Average Correlation Means / N(0.0, 1.0)



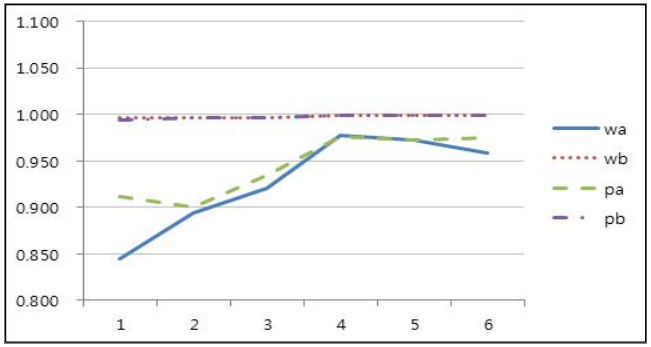
(b) Average Correlation SDs / N(0.0, 1.0)



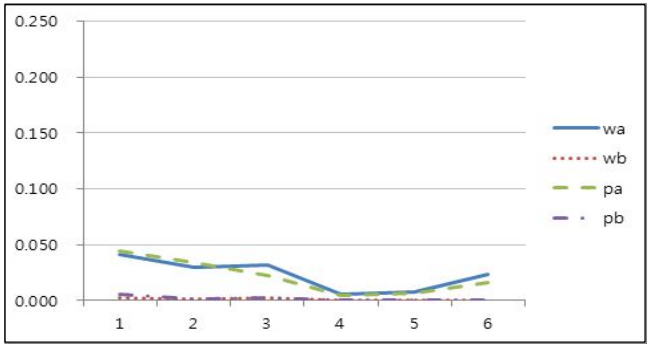
(c) Average Correlation Means / N(.25, 1.1^2)



(d) Average Correlation SDs / N(.25, 1.1^2)



(e) Average Correlation Means / N(.50, 1.2^2)



(f) Average Correlation SDs / N(.50, 1.2^2)

[Figure 5] Average Correlation Means and Standard Deviation

Note

1 : when N=500, #FI=10, 4 : when N=2000, #FI=10

2 : when N=500, #FI=20, 5 : when N=2000, #FI=20

3 : when N=500, #FI=40, 6 : when N=2000, #FI=40

wa: results of a-parameters when WinBUGS was used

wb: results of b-parameters when WinBUGS was used

pa: results of a-parameters when PARSCALE was used

pb: results of b-parameters when PARSCALE was used

Through Figure 3, first of all, it was clear that the two FIPC procedures have similar graphs when target group Q was one of N(0.0, 1.0), N(.25, 1.1^2) and N(.50, 1.2^2). The linking results of both a- and b- parameters have similar patterns across every condition. In addition, as sample sizes get larger, it shows that the correlations also become higher.

4.2 Root Mean Square Errors (RMSE)

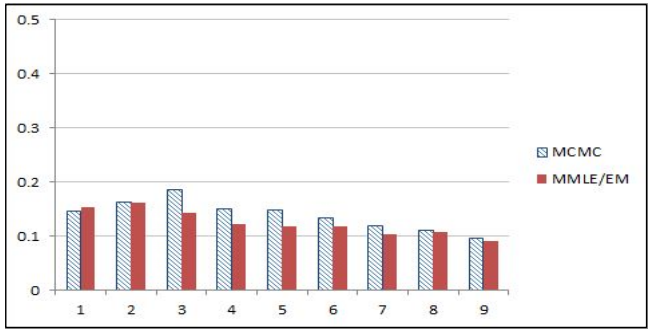
Table 4, Table 5 and Figure 4 present the result for RMSE values in all 18 simulation conditions. For sample size N=500 and in case of a-parameters, the average RMSE means and SDs across two FIPC procedures ranged between .091 and .186 and between .014 and .158, respectively. And in case of b-parameters, the RMSE means and SDs ranged between .107 and .211 and between .018 and .071, respectively. For sample size N=2000, the RMSE average means and SDs for a-parameters across two FIPC procedures ranged between .054 and .126 and between .005 and .020, respectively. And in case of b-parameters, the RMSE means and SDs ranged between .054 and .154 and between .005 and .025, respectively.

<Table 4> Average Means and Standard Deviations of RMSE values between Generating and Estimated Parameters of the Target Group (a-parameter)

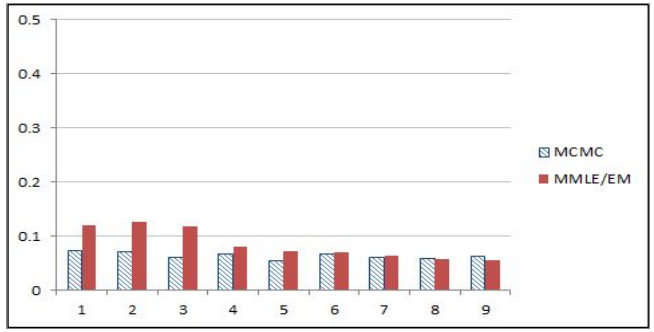
sample sizes	# common or fixed items	True Target Distributions	MCMC algorithm (WinBugs)		MMLE/EM algorithm (PARSCALE)	
			(PARSCALE)			
			mean	SD	mean	SD
N=500	#FI=10	N(0.0,1.0)	0.147	0.041	0.154	0.154
		N(.25,1.1^2)	0.164	0.069	0.161	0.023
		N(.50,1.2^2)	0.186	0.158	0.143	0.024
	#FI=20	N(0.0,1.0)	0.150	0.074	0.122	0.021
		N(.25,1.1^2)	0.149	0.045	0.118	0.027
		N(.50,1.2^2)	0.134	0.029	0.118	0.016
	#FI=40	N(0.0,1.0)	0.119	0.045	0.103	0.026
		N(.25,1.1^2)	0.111	0.017	0.107	0.021
		N(.50,1.2^2)	0.096	0.025	0.091	0.014
N=2000	#FI=10	N(0.0,1.0)	0.074	0.011	0.121	0.014
		N(.25,1.1^2)	0.071	0.014	0.126	0.014
		N(.50,1.2^2)	0.061	0.010	0.117	0.012
	#FI=20	N(0.0,1.0)	0.068	0.009	0.081	0.013
		N(.25,1.1^2)	0.054	0.005	0.073	0.013
		N(.50,1.2^2)	0.067	0.013	0.071	0.009
	#FI=40	N(0.0,1.0)	0.061	0.020	0.063	0.015
		N(.25,1.1^2)	0.058	0.013	0.057	0.012
		N(.50,1.2^2)	0.063	0.017	0.055	0.014

<Table 5> Average Means and Standard Deviations of RMSE values between Generating and Estimated Parameters of the Target Group (b-parameter)

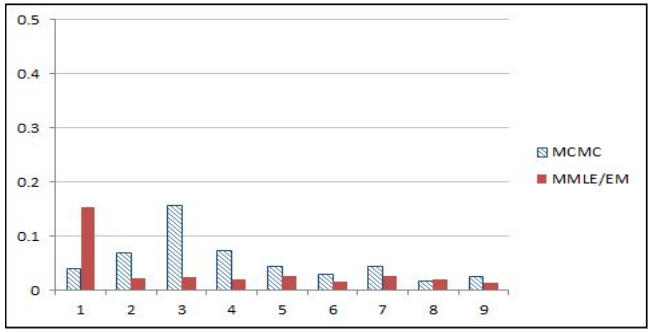
sample sizes	# common or fixed items	True Target Distributions	MCMC algorithm (WinBugs)		MMLE/EM algorithm (PARSCALE)	
			(PARSCALE)			
			mean	SD	mean	SD
N=500	#FI=10	N(0.0,1.0)	0.115	0.018	0.178	0.034
		N(.25,1.1^2)	0.108	0.019	0.164	0.035
		N(.50,1.2^2)	0.110	0.030	0.184	0.071
	#FI=20	N(0.0,1.0)	0.109	0.021	0.134	0.036
		N(.25,1.1^2)	0.118	0.018	0.123	0.022
		N(.50,1.2^2)	0.107	0.022	0.122	0.034
	#FI=40	N(0.0,1.0)	0.170	0.037	0.125	0.030
		N(.25,1.1^2)	0.211	0.048	0.151	0.050
		N(.50,1.2^2)	0.201	0.026	0.132	0.026
N=2000	#FI=10	N(0.0,1.0)	0.059	0.009	0.131	0.018
		N(.25,1.1^2)	0.055	0.005	0.127	0.021
		N(.50,1.2^2)	0.054	0.006	0.112	0.015
	#FI=20	N(0.0,1.0)	0.059	0.010	0.073	0.022
		N(.25,1.1^2)	0.054	0.011	0.074	0.018
		N(.50,1.2^2)	0.064	0.010	0.078	0.015
	#FI=40	N(0.0,1.0)	0.146	0.019	0.078	0.025
		N(.25,1.1^2)	0.154	0.016	0.063	0.023
		N(.50,1.2^2)	0.141	0.019	0.066	0.010



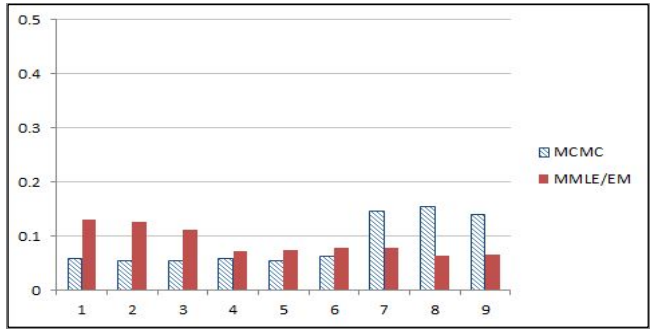
(a) Average The RMSE Means / N=500 / a-par



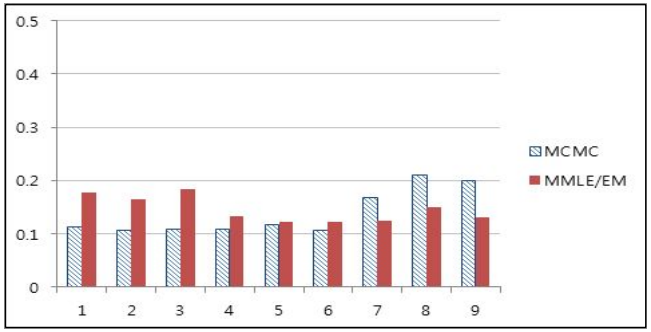
(b) Average The RMSE Means / N=2,000 / a-par



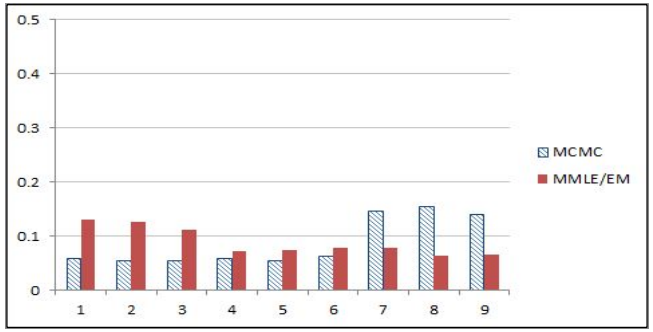
(c) Average The RMSE SDs / N=500 / a-par



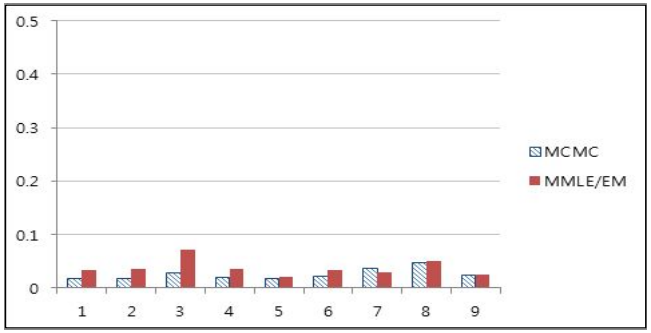
(d) Average The RMSE SDs / N=2,000 / a-par



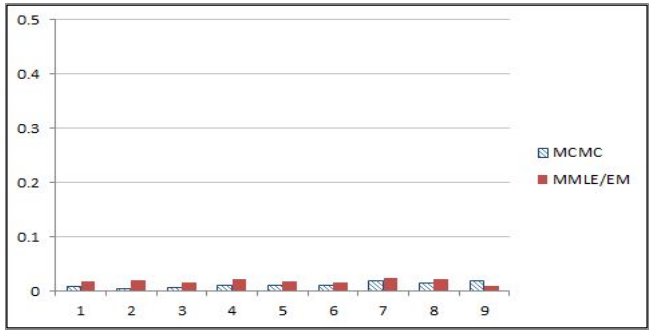
(e) Average The RMSE Means / N=500 / b-par



(f) Average The RMSE Means / N=2,000 / b-par



(g) Average The RMSE SDs / N=500 / b-par



(h) Average The RMSE SDs / N=2,000 / b-par

[Figure 6] Average the RMSE Means and Standard Deviations

Note.

1 : when #FI=10, N(0.0, 1.0)
2 : when #FI=10, N(.25, 1.1^2)
3 : when #FI=10, N(.50, 1.2^2)
4 : when #FI=20, N(0.0, 1.0)
5 : when #FI=20, N(.25, 1.1^2)
6 : when #FI=20, N(.50, 1.2^2)
7 : when #FI=40, N(0.0, 1.0)
8 : when #FI=40, N(.25, 1.1^2)
9 : when #FI=40, N(.50, 1.2^2)

Figure 4 showed good linking results for all conditions. In this figure, the total length of the bar represents the average RMSE means and SDs value. First, the two FIPC methods provided very similar RMSE results. Second, as the sample size increased from 500 to 2000, the accuracy of both linking methods appeared to get better. Third, when the sample size was 2,000, MCMC method conducted better than MMLE/EM algorithm in parameter linking through FIPC. Finally, for both FIPC methods, it was found that the linking accuracy was improved as the numbers of common or fixed items increase(#FI = 10, 20, or 40).

4.3 T-test

Table 6 presents the result for RMSE mean-difference tests between RMSE values calculated by two methods in the various simulation conditions. A positive mean-difference indicates the FIPC linking results using MCMC is better than those using MMLE/EM.

When the number of common or fixed item is 10(i.e., FI=10), the RMSE mean-differences of a-parameters and b-parameters are .020 and .066, respectively. Especially the result for t-test of the latter is significant under significant level of .05. For FI=20 conditions, the RMSE mean difference of a-parameters is -.007 and b-parameter is .034. And, these values are non-significant under the significant level .05. For FI=40, however, the RMSE mean differnece for both a- and b- parameters are significant. Then, MMLE/EM linking approach appeared to become more accurate than MCMC method, as the number of common or fixed item increased.

Also, for sample size N=500, the RMSE mean-difference of a-parameters is -.016 and b-parameter is .019. And for sample size=2000, the mean-differences of a-parameters and b-parameters are .021 and .002, respectively. Both results appeared to be statistically significant only for a-parameters. Finally, for true target distributions N(0.0, 1.0), N(.25, 1.1^2) and N(.50, 1.2^2), the RMSE mean-difference of a-parameters are .004, .006, -.003 and the values of b-parameteres are .029, .001, and .002, respectively. And all results of t-test are not significant.

<Taebl 6> RMSE mean difference between MCMC and MMLE linking methods

conditions		$\overline{d_{RMSE}}$ (SD)	
		a-parameter	b-parameter
# common or fixed Item (FI)	10	.020 (.080)	.066 (.029) *
	20	-.007 (.392)	.034 (.141)
	40	-.006 (.015) *	-.068 (.031) *
# examinee	500	-.016 (.065) *	.019 (.129)
	2000	.021 (.028) *	.002 (.063)
true target distribution	N(0.0,1.0)	.004 (.042)	.029 (.152)
	N(.25,1.1^2)	.006 (.041)	.001 (.063)
	N(.50,1.2^2)	-.003 (.072)	.002 (.063)

* $p < .05$

IV. Discussion

5.1 Summary

The main purpose of this study is to compare the results between MCMC and MMLE/EM algorithms when linking between two test forms is performed with the FIPC method. Through the simulation studies including various conditions, the item parameter linking results of two methods were compared as the sample size, the number of common items and ability distributions are changed; and it is confirmed whether the difference in the RMSE means of the two methods is significant.

For performing the study, simulation conditions included the number of examinees for both the base and target group(500 and 2000), the number of common items(10, 20, and 40) and ability distributions(N(0.0, 1.0), N(.25, 1.1^2), N(.50, 1.2^2)). After 10 data item sets were generated repeatedly in each condition, two equating methods were applied respectively and item parameters estimated. Under the various conditions, the correlation and the value of RMSE were calculated and compared to compare the performance of the two methods. Also, a t-test was conducted in order to determine whether the calculated RMSE values are significantly different.

In this study, the following research questions were proposed so

as to compare the performance of the two approaches according to a variety of conditions: (1) performance results of two equating methods were compared depending on the sample size, (2) performance results of two equating methods were compared depending on the number of common items, (3) performance results of two equating methods were compared depending on the ability distributions, (4) a check was performed to see whether the difference between the RMSE values of the two equating methods was significant.

The results of the research questions mentioned above are as follows.

First, when the sample size increased from 500 to 2,000, better performance is observed in both the MCMC and MMLE/EM algorithms. All methods produced higher correlations and lower RMSE values. These results mean that the accuracy of the two algorithms improved, as the sample size increased from 500 to 2,000.

Second, the performance of the two methods improved as the number of common items increased, and the results of the MCMC and MMLE/EM algorithms are similar. Also according to the results of the mean-difference tests, the FIPC using the MCMC algorithm appeared to be more accurate than the FIPC using the MMLE/EM algorithm when the number of fixed items is small(i.e., FI=10). On the other hand, the MMLE/EM method performed better than the MCMC algorithm when the number of common items was large.

Third, the results of the two approaches are similar in the three

ability distributions and the difference in ability distributions is negligible.

Finally, according to the results of the t-test, in order to verify whether the difference in the RMSE value in each condition is significant or not, there is a significant difference in the discrimination parameter when the sample size is 500 and 2,000. Also, when the number of fixed items is 40, the difference between the two methods is observed in both discrimination and difficulty parameters, and when the number of common items is 20, a significant difference in the difficulty parameter is found.

As a result, better results in the item parameter linking were produced in both algorithms as the sample size increased. Also, the superior results of item parameter linking were yielded in both MCMC and MMLE/EM algorithms as the number of common items increased and the results of the two methods were similar. Given the three ability distributions, because the calculated results of the two methods are alike in all conditions, the differences among the three ability distributions are not significant. Also, according to the result of the t-test in order to verify whether the value of RMSE is significant or not, there were significant differences in the discrimination parameters or difficulty parameters in some conditions. By determining that there is not a big difference between the results of the two calibration methods using the FIPC method when it is difficult to apply the MMLE/EM algorithm, it is expected that the linking between estimates of the item parameters can be easily

performed with the MCMC algorithm.

5.2 Conclusion and Discussion

This study was to compare two FIPC procedures: the MCMC-based FIPC conducted by the program WinBugs and the MMLE/EM-based FIPC conducted by the program PARSCALE. As expected, the two FIPC procedures performed similarly and showed good item-parameter linking performances. In all 18 simulation conditions, we could see that the linking results appeared to be pretty accurate and stable.

The effect of sample sizes on the linking results was observed through Figures 2 and 3. When the sample sizes increased from 500 to 2,000, the RMSE values decreased. These results meant that the accuracy of the two linking methods improved, as the sample sizes increased from 500 to 2,000. Relating the number of the common or fixed items, the linking performance seemed to be similar or be improved as the number of fixed items increased. According to the results of the mean-difference tests, the FIPC using the MCMC algorithm appeared to be more accurate than the FIPC using the MMLE/EM algorithm when the number of fixed items was small (i.e., FI=10). On the other hand, the MMLE/EM method performed better than the MCMC algorithm when the number of common items

was large.

Also, as the sample size is increased, both FIPC methods worked better for the item parameter linking. The results of the t-test indicated that the two FIPC methods performed similarly for b-parameter linking. But, there existed significant differences in the results of a-parameter linking. When the sample size is small, the FIPC using MMLE/EM seems to be more preferable. But, the FIPC using MCMC looks better when the sample size is 2,000. And the MCMC method seemed to perform better than MMLE/EM in almost every condition, although no related t-tests showed a statistical significance.

Overall, this study showed that both FIPC methods were useful tools for the purpose of item parameter linking. According to the results, although one method happened to provide better linking results than the other in some cases, the FIPC method based on the MCMC algorithm performed as well as the FIPC based on the MMLE/EM algorithm. As mentioned earlier, the MCMC method can be applied to more complex situations or studies. Then, the results of this study would be very useful for researchers and practitioners when they need to use the MCMC algorithm for estimating item parameters and linking the values to a base scale.

Also, the limitations of the present study are as follows. First of all, it needs to apply various models. The present study compared the two equating methods by using two parameter logistic models with the test consisting of dichotomous items. However, it is considered

that the present research should be conducted again with more multiple models in order to generalize the present study's results.

Second, it needs to apply to various simulation conditions. The present study analyzed the result of the experiment by setting up only three conditions of the simulation such as the number of examinees, fixed items and ability distribution of examinees; nevertheless, this research study requires more simulation conditions.

Lastly, further study should utilize more systematical approaches related with how to decide the prior distribution in the MCMC method. This study paralleled the results of two equating methods in accordance with the number of examinees, fixed items and ability distribution of the examinees in the same prior distribution, but further study is required to investigate using more and various prior distributions in the MCMC method.

Reference

Anfoff, W. H. (1987). Technical and practical issues in equating: A discussio of four papers. *Applied Psychological Measurement, 11*, 291-300.

Baker, F. and Kim, S.-H. (2004). *Item Response Theory: Parameter Estimation Techniques* 2nd Ed. New York, NY: Marcel Dekker, Inc.

Baldwin, S. G., Baldwin, P., & Nering, M. L. (2007). *A comparison of IRT equating methods on recovering item parameters and growth in mixed-format tests*. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

Beichl, I., & Sullivan, F. (2000). The Metropolis algorithm. *Computing in Science & Engineering, 2:1*, 65-69.

Bergman, N. (1999). Recursive Bayesian estimation: Navigation and tracking applications. Ph.D. Thesis, Department of Electrical Engineering, Linkoping University, Sweden.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: An application of an EM algorithm. *Psychometrika, 46*, 443-459.

Brennan, R. L. & Kolen, M. J. (1987). Some practical issues in equating.. *Applied Psychological Measurement, 11*, 279-290.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood for incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B, 39*, 1-38.

Dorans, N. J. (1990). Equating methods and sampling designs. *Applied Measurement in Education, 3*, 3-17.

Drasgow, F. (1989). An evaluation of marginal maximum likelihood estimation for the two-parameter logistic model. *Applied Psychological Measurement, 13*, 77-90.

Dyer, M., Frieze, A., & Kannan, R. (1991). A random polynomial-time algorithm for approximating the volume of convex bodies. *Journal of the ACM, 1:38*, 1-17.

Hambleton, R. K. (1983). *Application of IRT.* Vancouver: Educational Recearch Institute of British Columbia.

Han, J. A., (2011). Influences of conditions for mixed format tests on the performance of IRT equating under the common item nonequivalent design. M.A. Dissertation, Yonsei University.

Han, K. T., (2010) WinGen3: *Windows software that generates IRT parameters and item response* [Computer program]. Amherst, MA: University of Massachusetts, Center for Educational Assessment.

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

Harris, D. J. (1993). Practical issues in equating. Paper presented at the annual meeting of the American Educational Research Association, Atlanta.

Harwell, M. R., Baker, F. B., & Zwarts, M. (1988). Item parameter estimation via marginal maximum likelihood and an EM algorithm. *Journal of Educational Statistics, 13,* 243-271.

Jerrum, M., & Sinclair, A. (1996). *The Markov chain Monte Carlo method: an approach to approximate counting and integration.* In D. S. Hochbaum (Ed.), Approximation algorithms for NP-hard problems (pp. 482 - -519). PWS Publishing.

Kang, T., & Petersen, N. S. (2012). Linking item parameters to a base scale. *Asia Pacific Education Review, 13(2),* 311-321

Keller, R. R., Keller, L. A., & Baldwin, S. (2007). *The effect of changing equating methods on monitoring growth in mixed-format tests.* Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.

Kim, J., Frisbie, D. A., Kolen, M. J., & Kim, D.-I. (2007). *A comparison of calibration methods and proficiency estimators for creating IRT vertical scales.* Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.

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

Kim, S.-H., & Cohen, A. S. (1998). A comparison of linking and concurrent calibration under item response theory. *Applied Psychological Measurement, 22*, 131-143.

Kim, S.-H., & Cohen, A. S. (2002). A comparison of linking and concurrent calibration under the graded response model. *Applied Psychological Measurement, 26*, 25-41.

Kim, Y. J. (2010). An investigation of the Effect of the Anchor Test Length and the Non-equivalency of Equating Group on Equating, M.A. Dissertation, Yonsei University.

Kolen, M. J. (1981). Comparison of traditional and IRT methods for equating tests. *Journal of Educational Measurement, 18*, 1-11.

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. (submitted). Comparison of three IRT linking procedures in the random groups equating design. *Applied Measurement in Education*.

Li, Y. H., Griffith, W. D., & Tam, H. P. (1997). *Equating multiple tests via an IRT linking design: Utilizing a single set of anchor items with fixed common item parameters during the calibration process*. Paper presented at the annual meeting of the Psychometric Society, Knoxville, TN.

Lim, E. J. (2011), Comparison of small-sample equating methods for mixed-format tests in a NEAT design, M.A. Dissertation, Yonsei University.

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

Mislevy, R. J. (1992). *Linking educational assessment: Conceps, issues, method, and prospects*. Princeton, NJ: ETS Policy Information Center.

Mislevy, R. J., & Stocking, M. L. (1989). A consumer''s guide to LOGIST and BILOG. *Applied Psychological Measurement, 13*, 57-75.

Muraki, E., & Bock, R. D. PARSCALE: *IRT based item analysis and test scoring for rating-scale data* [Computer program]. Chicago, IL: Scientific Software International.

Muraki, E., Hombo, C. M., & Lee, Y.-W. (2000). Equating and liking of performance assessments. *Applied Psychological Measurement, 24*, 325-337.

Nam, H. W. (2001). *Methods of Test Equating*. Seoul, Kyoyookbook.

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.

Pak, S. H. (2010), A Comparison of Kernel and traditional equating methods under the non-equivalent groups with anchor test design, M.A. Dissertation, Yonsei University.

Qualls, A. L., & Ansley, T. N. (1985). *A comparison of item and ability parameter estimates derived from LOGIST and B.ILOG Paper presented at the meeting of the National Council on Measurement in Education,* Chicago IL, U.S.A.

Scorupski, W. P., Jodoin, M. G., Keller, L. A., & Swaminathan, H. (2003). *An evaluation of item response theory equating procedures for capturing growth with tests composed of dichotomously scored items*. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.

Seong, Tae-je. (1990). Sensitivity of marginal maximum likelihood estimation of item and ability parameters to the characteristics of the prior ability distributions. *Applied Psychological Measurement, 14*, 299-311.

Seong, T. J. (2001). *Understanding and Application of Item Response Theory.* Seoul, Kyoyookbook.

Sinharay, S. & Holland, P. W. (2010). A new approach to comparing several equating methods in the context of the NEAT design. *Journal of Educational Measurement, 47(3)*, 261-285.

Song, M. Y., Nam, M. W., Kang, T. H. & Kim, C. I. (2011). Vertical Scaling for National Assessment of Educational Achievement to measure individual students' growth. *Korea Institute for Curriculum and Evaluation.* RRE 2011-6-1.

Swaminathan, H., & Gifford, J. A. (1985). Bayesian estimation in the two-parameter logistic model. *Psychometrika, 50*, 349-364.

Vale, C. D. (1986). Linking item parameters onto a common scale. *Applied Psychological Measurement, 10*, 333-344.

Yen, M. Y. (1987). A comparison of the efficiency and accuracy of BILOG and LOGIST. *Psychometrika, 52*, 275-291.

Yi, H. (2009). Evaluating the performance of non-equivalent groups anchor test equating under various conditions of anchor test construction. *Journal of Educational Evaluating, 22*, 847-869.

ABSTRACT

A Comparison of MCMC and MMLE/EM Algorithms for Fixed Item Parameter Calibration

So Hee, Kim

Departmet of Education

The graduate school of Sungshin Women's University

The main purpose of this study is to compare if there are any differences of performance between MCMC algorithm and MMLE/EM algorithm through a simulation study, when item parameters from different test forms are linked with each FIPC method. By comparing the performance of two algorithms, it will be useful to check the utility of MCMC algorithm for FIPC because the application can be easily extended to much more complicated situations where the MMLE/EM algorithm is very difficult to apply.

The specific objectives of this study comparing the MCMC and MMLE/EM algorithms were as follows:

1. To compare the results of the two methods as sample size changes

2. To compare the results of the two methods as the number of common or fixed item changes

3. To compare the results of the two methods as true target distribution changes

4. To investigate whether the difference in RMSE values produced by the two methods is significant

To answer these questions, simulation studies were conducted. Simulation conditions were considered for this study including the number of examinees(N=500, 2000), the number of common items(CI=10, 20, 40) and ability distribution(N(0.0, 1.0), N(.25, 1.1^2), N(.50, 1.2^2)). Two linking methods were applied after 10 data sets were generated repeatedly in each condition and then item parameters were estimated respectively. To compare the performance of the two algorithms, the Pearson's product moment correlations and RMSE were computed and compared. And then, t-test was conducted in order to check whether the difference of RMSE values calculated is significant.

And the results of this study were as follows:

1. When the sample size increased from 500 to 2,000, better performance is observed in both the MCMC and the MMLE/EM

algorithms and both methods produced higher correlation and lower RMSE values.

2. The performance of the two methods is better as the number of common items increased and the results of the MCMC and MMLE/EM algorithms are similar.

3. The results of two approaches are similar in three ability distributions and the difference in ability distribution is negligible.

4. According to the results of t-test in order to verify whether the difference in RMSE value in each condition is significant or not, there is a significant difference in discrimination parameter and when the number of fixed item is 40, the difference between two methods is observed in both discrimination and difficulty parameters.

Through the above results, this study showed that both algorithms were useful tool for the purpose of item parameter linking. In addition, a further study including a variety of models, simulation conditions, prior distributions need to be performed.

Keywords : item response theory, linking, FIPC, MMLE/EM algorithms, MCMC

APPENDIX

APPENDIX 1 : PARSCALE CODE for FIPC

APPENDIX 2 : WinBUGS CODE for FIPC

APPENDIX 3 : Result of correlation for each condition

APPENDIX 4 : Result of RMSE for each condition

APPENDIX 1 : PARSCALE CODE for FIPC

```
-------------------------------------------------------------
FIPC Study Calibration
Generating item parameters are from 2005 AAP Math 65 or 66A
>COMMENT
>FILE    DFNAME='new101.dat', IFNAME='PSLold(common10_101).prm', SAVE;
>SAVE   PARM='fix101.par';
>INPUT NIDCH=4, NTOTAL=50, NTEST=1, LENGTH=50, NFMT=1;
(4A1, T1, 50A1)
>TEST     TNAME=Math, ITEM=(01(1)50), NBLOCK=50,
INAMES=(
        CO01, CO02, CO03, CO04, CO05, CO06, CO07, CO08, CO09, CO10,
          NE01, NE02, NE03, NE04, NE05, NE06, NE07, NE08, NE09, NE10,
          NE11, NE12, NE13, NE14, NE15, NE16, NE17, NE18, NE19, NE20,
          NE21, NE22, NE23, NE24, NE25, NE26, NE27, NE28, NE29, NE30,
          NE31, NE32, NE33, NE34, NE35, NE36, NE37, NE38, NE39, NE40);
>BLOCK1    BNAME=COMMON,  NITEM=1,  NCAT=2,  ORI=(0,1),  MOD=(1,2),
REP=10, SKIP;
>BLOCK2    BNAME=COMMON,  NITEM=1,  NCAT=2,  ORI=(0,1),  MOD=(1,2),
REP=40;
>CALIB    PARTIAL, LOGISTIC, SCALE=1.7, NQPT=11, CYCLE=(3000,1,1,1,1),
           FREE=(NOADJUST, NOADJUST), POSTERIOR, NEWTON=0,
           CRIT=0.001, ITEMFIT=10, SPRIOR, GPRIOR;
>SCORE ;
-------------------------------------------------------------
```

APPENDIX 2 : WinBUGS CODE for FIPC

```
model {
# m3_ : We are using 2 Parameter model in this Winbugs run.
# This data are an initial calibration sample
  for (j in 1:N) {
   for (k in 1:T) {
          r[j,k]<-resp[j,k]
}}

# 2PL model
  for (j in 1:N) {
      for (k in 1:T) {
          tt[j,k]<- exp(-1.7*a[k]*(theta[j] - b[k]))
          p[j,k]<- 1 /(1 + tt[j,k])
          r[j,k]~dbern(p[j,k])
          }
       theta[j] ~ dnorm(mu,isig)
          }

#   Priors
mu~dnorm(0,.0001)
isig~dchisqr(.5)
for (k in 11:T) {
    a[k]~dlnorm(0.,1.)
    b[k]~dnorm(0.,1.)
    }
```

```
 a[1]<-  1.22564  ; b[1]<-    .80877  ;
 a[2]<-  1.39454  ; b[2]<-    .29189  ;
 a[3]<-   .95372  ; b[3]<-  -1.03176  ;
 a[4]<-   .89319  ; b[4]<-   -.42274  ;
 a[5]<-   .80851  ; b[5]<-  -1.69194  ;
 a[6]<-  1.00819  ; b[6]<-   -.57175  ;
 a[7]<-  1.39760  ; b[7]<-   -.08690  ;
 a[8]<-  1.27927  ; b[8]<-  -1.35591  ;
 a[9]<-  1.37410  ; b[9]<-    .17912  ;
 a[10]<- 1.18630  ; b[10]<-   .01664  ;
}
------------------------------------------------------------
```

APPENDIX 3 : Result of correlation for each condition

condition	MCMC algorithm		MMLE/EM algorithm	
	a-parameter	b-parameter	a-parameter	b-parameter

111
N=500
#FI=10
N(0.0,1.0)
	0.885	0.997	0.847	0.996

	0.890	0.997	0.887	0.997
	0.903	0.996	0.909	0.996
	0.911	0.994	0.892	0.993
	0.881	0.995	0.863	0.995
	0.947	0.995	0.934	0.996
	0.917	0.996	0.906	0.996
	0.917	0.994	0.923	0.994
	0.942	0.995	0.934	0.995
	0.845	0.997	0.898	0.996
MEAN	0.904	0.995	0.899	0.995
SD	0.030	0.001	0.029	0.001

112
N=500
#FI=10
N(.25,1.1^2)
	0.860	0.996	0.874	0.996

	0.872	0.998	0.920	0.998
	0.908	0.996	0.890	0.995
	0.866	0.996	0.859	0.995
	0.909	0.996	0.888	0.996
	0.907	0.998	0.906	0.998
	0.712	0.996	0.838	0.996
	0.879	0.995	0.884	0.995
	0.926	0.996	0.927	0.996
	0.906	0.994	0.884	0.994
MEAN	0.874	0.996	0.887	0.996
SD	0.061	0.001	0.027	0.001

113
N=500
#FI=10
N(.50,1.2^2)
	0.915	0.996	0.912	0.996

	0.859	0.989	0.796	0.976
	0.837	0.995	0.915	0.995
	0.920	0.997	0.903	0.996
	0.253	0.998	0.928	0.997
	0.933	0.998	0.947	0.997
	0.941	0.997	0.903	0.994
	0.926	0.997	0.939	0.997
	0.947	0.997	0.943	0.996
	0.923	0.997	0.938	0.998
MEAN	0.845	0.996	0.912	0.994
SD	0.211	0.003	0.044	0.006

condition	MCMC algorithm		MMLE/EM algorithm	
	a-parameter	b-parameter	a-parameter	b-parameter

121
N=500
#FI=20
N(0.0,1.0)
	0.861	0.995	0.882	0.996

	0.733	0.997	0.900	0.995
	0.908	0.998	0.903	0.998
	0.884	0.995	0.881	0.996
	0.910	0.997	0.918	0.997
	0.893	0.995	0.918	0.994
	0.934	0.998	0.910	0.998
	0.880	0.992	0.843	0.990
	0.970	0.997	0.960	0.997
	0.914	0.997	0.888	0.997
MEAN	0.889	0.996	0.900	0.996
SD	0.063	0.002	0.030	0.002

122
N=500
#FI=20
N(.25,1.1^2)
	0.918	0.996	0.901	0.996

	0.913	0.997	0.923	0.997
	0.920	0.995	0.940	0.996
	0.848	0.994	0.877	0.994
	0.886	0.996	0.862	0.996
	0.900	0.997	0.904	0.997
	0.793	0.997	0.800	0.997
	0.941	0.996	0.931	0.997
	0.889	0.994	0.894	0.994
	0.887	0.996	0.916	0.996
MEAN	0.889	0.996	0.895	0.996
SD	0.042	0.001	0.041	0.001

123
N=500
#FI=20
N(.50,1.2^2)
	0.913	0.998	0.916	0.999

	0.872	0.996	0.907	0.996
	0.939	0.996	0.927	0.996
	0.857	0.996	0.906	0.997
	0.884	0.998	0.912	0.998
	0.847	0.997	0.816	0.997
	0.885	0.997	0.876	0.997
	0.921	0.994	0.935	0.994
	0.919	0.993	0.897	0.993
	0.902	0.996	0.907	0.995
MEAN	0.894	0.996	0.900	0.996
SD	0.030	0.002	0.034	0.002
131	0.921	0.998	0.950	0.999
	0.877	0.996	0.874	0.995

condition	MCMC algorithm		MMLE/EM algorithm	
	a-parameter	b-parameter	a-parameter	b-parameter
N=500 #FI=40 N(0.0,1.0)	0.827	0.996	0.841	0.996
	0.897	0.996	0.916	0.996
	0.986	0.997	0.974	0.996
	0.968	0.999	0.942	0.999
	0.672	0.998	0.800	0.998
	0.901	0.997	0.923	0.996
	0.916	0.997	0.914	0.997
	0.846	0.995	0.910	0.995
MEAN	0.881	0.997	0.904	0.997
SD	0.088	0.001	0.052	0.001
132 N=500 #FI=40 N(.25,1.1^2)	0.751	0.997	0.808	0.997
	0.899	0.996	0.902	0.996
	0.940	0.999	0.912	0.999
	0.895	0.997	0.895	0.997
	0.892	0.996	0.890	0.994
	0.902	0.992	0.910	0.993
	0.861	0.996	0.843	0.996
	0.929	0.998	0.969	0.999
	0.969	0.995	0.979	0.996
	0.931	0.996	0.937	0.996
MEAN	0.897	0.996	0.905	0.996
SD	0.060	0.002	0.052	0.002
133 N=500 #FI=40 N(.50,1.2^2)	0.949	0.998	0.949	0.998
	0.892	0.998	0.901	0.998
	0.953	0.995	0.946	0.996
	0.933	0.997	0.956	0.997
	0.904	0.997	0.909	0.998
	0.874	0.990	0.918	0.991
	0.930	0.992	0.932	0.991
	0.974	0.998	0.969	0.997
	0.893	0.996	0.919	0.996
	0.909	0.996	0.936	0.996
MEAN	0.921	0.996	0.934	0.996
SD	0.032	0.003	0.022	0.003
211	0.968	0.999	0.966	0.999
	0.971	0.998	0.975	0.998
	0.975	0.999	0.972	0.999

condition	MCMC algorithm		MMLE/EM algorithm	
	a-parameter	b-parameter	a-parameter	b-parameter
N=2,000 #FI=10 N(0.0,1.0)	0.971	0.999	0.969	0.999
	0.950	0.999	0.944	0.999
	0.988	0.999	0.981	0.999
	0.968	0.998	0.965	0.998
	0.967	0.999	0.965	0.998
	0.981	0.999	0.981	0.999
	0.979	0.999	0.976	0.999
MEAN	0.972	0.999	0.969	0.999
SD	0.010	0.000	0.011	0.000
212 N=2,000 #FI=10 N(.25,1.1^2)	0.975	0.999	0.973	0.999
	0.957	0.999	0.955	0.999
	0.977	0.999	0.976	0.999
	0.960	0.999	0.959	0.999
	0.964	0.999	0.971	0.999
	0.977	0.999	0.974	0.999
	0.972	0.999	0.968	0.999
	0.978	0.999	0.973	0.999
	0.976	0.999	0.975	0.999
	0.976	0.999	0.971	0.999
MEAN	0.971	0.999	0.969	0.999
SD	0.008	0.000	0.007	0.000
213 N=2,000 #FI=10 N(.50,1.2^2)	0.980	0.999	0.978	0.999
	0.972	0.999	0.966	0.999
	0.977	0.999	0.980	0.999
	0.984	0.999	0.980	0.999
	0.971	0.999	0.974	0.999
	0.974	0.999	0.974	0.999
	0.987	0.999	0.982	0.999
	0.978	0.999	0.975	0.999
	0.966	0.998	0.968	0.998
	0.977	0.999	0.977	0.999
MEAN	0.977	0.999	0.976	0.999
SD	0.006	0.000	0.005	0.000
221	0.963	0.999	0.960	0.999
	0.982	1.000	0.980	1.000

condition	MCMC algorithm		MMLE/EM algorithm	
	a-parameter	b-parameter	a-parameter	b-parameter
N=2,000 #FI=20 N(0.0,1.0)	0.962	0.999	0.961	0.999
	0.972	0.999	0.976	0.999
	0.955	0.999	0.962	0.999
	0.982	0.998	0.979	0.998
	0.970	0.999	0.965	0.999
	0.957	0.999	0.956	0.999
	0.971	0.999	0.964	0.998
	0.964	0.999	0.961	0.999
MEAN	0.968	0.999	0.966	0.999
SD	0.009	0.000	0.009	0.000
222 N=2,000 #FI=20 N(.25,1.1^2)	0.988	0.999	0.985	0.999
	0.982	1.000	0.980	1.000
	0.971	0.999	0.975	0.999
	0.981	0.999	0.977	0.999
	0.967	0.999	0.965	0.999
	0.978	0.999	0.972	0.999
	0.986	0.999	0.986	0.999
	0.983	0.999	0.985	0.999
	0.985	0.999	0.982	0.999
	0.982	0.999	0.982	0.999
MEAN	0.980	0.999	0.979	0.999
SD	0.007	0.000	0.007	0.000
223 N=2,000 #FI=20 N(.50,1.2^2)	0.967	0.999	0.966	0.999
	0.973	0.998	0.972	0.998
	0.966	0.999	0.966	0.999
	0.957	0.999	0.963	0.999
	0.978	0.999	0.979	0.999
	0.968	0.999	0.969	0.999
	0.985	0.999	0.982	0.999
	0.977	0.998	0.980	0.998
	0.971	0.998	0.972	0.998
	0.975	0.998	0.974	0.998
MEAN	0.972	0.999	0.972	0.999
SD	0.008	0.000	0.007	0.000
231	0.936	0.998	0.962	0.998
	0.992	1.000	0.987	1.000

condition	MCMC algorithm		MMLE/EM algorithm	
	a-parameter	b-parameter	a-parameter	b-parameter
N=2,000 #FI=40 N(0.0,1.0)	0.960	0.999	0.965	0.999
	0.985	0.999	0.989	0.999
	0.968	0.998	0.974	0.998
	0.947	0.999	0.950	1.000
	0.988	0.998	0.983	0.998
	0.975	0.999	0.975	0.999
	0.970	0.999	0.985	0.999
	0.947	0.998	0.959	0.999
MEAN	0.967	0.999	0.973	0.999
SD	0.019	0.001	0.013	0.001
232 N=2,000 #FI=40 N(.25,1.1^2)	0.970	0.999	0.969	0.999
	0.975	0.999	0.987	1.000
	0.995	1.000	0.994	1.000
	0.979	0.999	0.996	1.000
	0.958	0.999	0.967	0.999
	0.962	1.000	0.985	1.000
	0.990	0.999	0.991	1.000
	0.955	0.998	0.964	0.998
	0.967	0.999	0.979	0.999
	0.979	0.999	0.992	1.000
MEAN	0.973	0.999	0.982	0.999
SD	0.013	0.001	0.012	0.001
233 N=2,000 #FI=40 N(.50,1.2^2)	0.985	0.999	0.977	0.999
	0.915	0.999	0.938	0.999
	0.979	0.999	0.981	0.999
	0.954	0.998	0.985	0.999
	0.941	0.999	0.961	0.999
	0.979	0.999	0.983	0.999
	0.977	1.000	0.990	1.000
	0.934	0.998	0.969	0.998
	0.957	1.000	0.981	1.000
	0.971	0.999	0.985	0.999
MEAN	0.959	0.999	0.975	0.999
SD	0.023	0.000	0.016	0.000

APPENDIX 4 : Result of RMSE for each condition

condition	MCMC algorithm		MMLE/EM algorithm	
	a-parameter	b-parameter	a-parameter	b-parameter

111
N=500
#FI=10
N(0.0,1.0)
	0.146	0.105	0.188	0.210

	0.210	0.097	0.151	0.165
	0.125	0.103	0.169	0.183
	0.153	0.123	0.162	0.195
	0.142	0.114	0.149	0.149
	0.092	0.112	0.146	0.151
	0.119	0.108	0.140	0.181
	0.136	0.154	0.170	0.249
	0.128	0.133	0.118	0.151
	0.224	0.097	0.149	0.143
MEAN	0.147	0.115	0.154	0.178
SD	0.041	0.018	0.019	0.034

112
N=500
#FI=10
N(.25,1.1^2)
	0.193	0.108	0.158	0.181

	0.174	0.078	0.131	0.119
	0.144	0.115	0.150	0.136
	0.143	0.109	0.158	0.141
	0.119	0.112	0.188	0.212
	0.149	0.077	0.140	0.132
	0.346	0.104	0.188	0.173
	0.135	0.123	0.185	0.211
	0.106	0.110	0.131	0.138
	0.129	0.142	0.180	0.198
MEAN	0.164	0.108	0.161	0.164
SD	0.069	0.019	0.023	0.035

113
N=500
#FI=10
N(.50,1.2^2)
	0.124	0.112	0.136	0.130

	0.163	0.185	0.202	0.373
	0.223	0.131	0.159	0.197
	0.133	0.100	0.143	0.154
	0.623	0.092	0.141	0.183
	0.100	0.080	0.145	0.145
	0.096	0.110	0.140	0.208
	0.161	0.100	0.110	0.129
	0.094	0.107	0.139	0.170
	0.145	0.084	0.120	0.156
MEAN	0.186	0.110	0.143	0.184
SD	0.158	0.030	0.024	0.071

condition	MCMC algorithm		MMLE/EM algorithm	
	a-parameter	b-parameter	a-parameter	b-parameter

121
N=500
#FI=20
N(0.0,1.0)
	0.228	0.121	0.149	0.129

	0.325	0.100	0.123	0.167
	0.129	0.105	0.119	0.134
	0.119	0.122	0.143	0.142
	0.120	0.094	0.110	0.121
	0.157	0.120	0.114	0.136
	0.129	0.081	0.117	0.086
	0.127	0.155	0.143	0.212
	0.068	0.093	0.077	0.095
	0.104	0.101	0.124	0.117
MEAN	0.150	0.109	0.122	0.134
SD	0.074	0.021	0.021	0.036

122
N=500
#FI=20
N(.25,1.1^2)
	0.105	0.103	0.108	0.113

	0.121	0.092	0.096	0.100
	0.148	0.117	0.101	0.121
	0.193	0.149	0.134	0.168
	0.124	0.118	0.134	0.141
	0.178	0.134	0.111	0.101
	0.237	0.094	0.181	0.111
	0.096	0.112	0.090	0.114
	0.117	0.134	0.121	0.146
	0.172	0.127	0.105	0.121
MEAN	0.149	0.118	0.118	0.123
SD	0.045	0.018	0.027	0.022

123
N=500
#FI=20
N(.50,1.2^2)
	0.113	0.078	0.099	0.060

	0.152	0.100	0.110	0.131
	0.091	0.110	0.122	0.132
	0.200	0.108	0.109	0.098
	0.137	0.089	0.104	0.075
	0.140	0.088	0.143	0.133
	0.122	0.104	0.146	0.140
	0.143	0.138	0.113	0.162
	0.113	0.150	0.121	0.159
	0.129	0.107	0.111	0.127
MEAN	0.134	0.107	0.118	0.122
SD	0.029	0.022	0.016	0.034
131	0.144	0.190	0.113	0.158
	0.126	0.186	0.110	0.133

condition	MCMC algorithm		MMLE/EM algorithm	
	a-parameter	b-parameter	a-parameter	b-parameter
N=500 #FI=40 N(0.0,1.0)	0.181	0.177	0.143	0.115
	0.131	0.189	0.112	0.152
	0.037	0.213	0.057	0.127
	0.059	0.137	0.075	0.057
	0.171	0.100	0.137	0.126
	0.097	0.120	0.089	0.154
	0.118	0.189	0.100	0.101
	0.123	0.195	0.094	0.129
MEAN	0.119	0.170	0.103	0.125
SD	0.045	0.037	0.026	0.030
132 N=500 #FI=40 N(.25,1.1^2)	0.151	0.162	0.132	0.115
	0.099	0.203	0.098	0.164
	0.100	0.154	0.111	0.054
	0.120	0.216	0.107	0.153
	0.110	0.235	0.116	0.181
	0.106	0.220	0.115	0.189
	0.120	0.249	0.134	0.171
	0.100	0.140	0.069	0.115
	0.089	0.237	0.076	0.136
	0.117	0.297	0.110	0.236
MEAN	0.111	0.211	0.107	0.151
SD	0.017	0.048	0.021	0.050
133 N=500 #FI=40 N(.50,1.2^2)	0.078	0.211	0.080	0.119
	0.111	0.222	0.112	0.112
	0.068	0.198	0.076	0.123
	0.105	0.180	0.090	0.105
	0.099	0.211	0.099	0.122
	0.147	0.228	0.113	0.176
	0.087	0.207	0.088	0.181
	0.060	0.231	0.074	0.141
	0.103	0.158	0.094	0.122
	0.104	0.160	0.082	0.119
MEAN	0.096	0.201	0.091	0.132
SD	0.025	0.026	0.014	0.026
211	0.072	0.056	0.128	0.122
	0.075	0.073	0.108	0.122
	0.064	0.057	0.141	0.168

condition	MCMC algorithm		MMLE/EM algorithm	
	a-parameter	b-parameter	a-parameter	b-parameter
N=2,000 #FI=10 N(0.0,1.0)	0.073	0.045	0.120	0.132
	0.094	0.063	0.130	0.137
	0.064	0.055	0.100	0.114
	0.070	0.067	0.143	0.151
	0.095	0.072	0.110	0.108
	0.066	0.051	0.116	0.131
	0.067	0.052	0.115	0.123
MEAN	0.074	0.059	0.121	0.131
SD	0.011	0.009	0.014	0.018
212 N=2,000 #FI=10 N(.25,1.1^2)	0.072	0.063	0.115	0.116
	0.086	0.063	0.112	0.096
	0.064	0.047	0.109	0.095
	0.101	0.049	0.116	0.126
	0.074	0.056	0.137	0.123
	0.062	0.055	0.116	0.121
	0.065	0.056	0.141	0.146
	0.059	0.052	0.130	0.144
	0.062	0.055	0.138	0.158
	0.060	0.049	0.147	0.142
MEAN	0.071	0.055	0.126	0.127
SD	0.014	0.005	0.014	0.021
213 N=2,000 #FI=10 N(.50,1.2^2)	0.057	0.051	0.116	0.111
	0.065	0.051	0.135	0.134
	0.061	0.055	0.105	0.092
	0.049	0.048	0.135	0.132
	0.070	0.053	0.111	0.099
	0.065	0.051	0.120	0.117
	0.045	0.051	0.120	0.110
	0.058	0.057	0.109	0.098
	0.079	0.067	0.096	0.105
	0.061	0.054	0.120	0.126
MEAN	0.061	0.054	0.117	0.112
SD	0.010	0.005	0.012	0.015
221	0.072	0.055	0.080	0.076
	0.057	0.043	0.052	0.040

condition	MCMC algorithm		MMLE/EM algorithm	
	a-parameter	b-parameter	a-parameter	b-parameter
N=2,000 #FI=20 N(0.0,1.0)	0.079	0.057	0.074	0.064
	0.060	0.049	0.087	0.070
	0.074	0.060	0.088	0.069
	0.054	0.077	0.073	0.115
	0.064	0.066	0.101	0.105
	0.079	0.056	0.088	0.053
	0.074	0.068	0.078	0.073
	0.066	0.058	0.090	0.069
MEAN	0.068	0.059	0.081	0.073
SD	0.009	0.010	0.013	0.022
222 N=2,000 #FI=20 N(.25,1.1^2)	0.051	0.060	0.049	0.064
	0.054	0.038	0.062	0.045
	0.059	0.051	0.079	0.077
	0.055	0.073	0.093	0.112
	0.064	0.067	0.081	0.070
	0.053	0.049	0.067	0.056
	0.050	0.041	0.069	0.076
	0.054	0.054	0.088	0.087
	0.047	0.053	0.077	0.082
	0.053	0.057	0.066	0.074
MEAN	0.054	0.054	0.073	0.074
SD	0.005	0.011	0.013	0.018
223 N=2,000 #FI=20 N(.50,1.2^2)	0.064	0.058	0.075	0.058
	0.063	0.067	0.070	0.082
	0.066	0.073	0.077	0.097
	0.076	0.055	0.071	0.067
	0.072	0.045	0.073	0.069
	0.081	0.064	0.082	0.086
	0.045	0.056	0.053	0.059
	0.056	0.070	0.061	0.075
	0.091	0.080	0.066	0.097
	0.055	0.068	0.083	0.094
MEAN	0.067	0.064	0.071	0.078
SD	0.013	0.010	0.009	0.015
231	0.084	0.130	0.063	0.088
	0.031	0.156	0.046	0.042

condition	MCMC algorithm		MMLE/EM algorithm	
	a-parameter	b-parameter	a-parameter	b-parameter
N=2,000 #FI=40 N(0.0,1.0)	0.066	0.138	0.066	0.084
	0.038	0.148	0.048	0.057
	0.061	0.178	0.054	0.112
	0.087	0.127	0.075	0.043
	0.036	0.170	0.051	0.097
	0.070	0.153	0.088	0.086
	0.056	0.124	0.057	0.063
	0.077	0.133	0.083	0.107
MEAN	0.061	0.146	0.063	0.078
SD	0.020	0.019	0.015	0.025
232 N=2,000 #FI=40 N(.25,1.1^2)	0.050	0.137	0.037	0.030
	0.033	0.159	0.037	0.052
	0.071	0.146	0.069	0.069
	0.068	0.166	0.072	0.072
	0.064	0.152	0.056	0.036
	0.042	0.178	0.058	0.085
	0.069	0.181	0.064	0.095
	0.056	0.142	0.057	0.070
	0.066	0.139	0.066	0.086
	0.060	0.142	0.050	0.038
MEAN	0.058	0.154	0.057	0.063
SD	0.013	0.016	0.012	0.023
233 N=2,000 #FI=40 N(.50,1.2^2)	0.040	0.161	0.050	0.069
	0.092	0.164	0.086	0.064
	0.047	0.126	0.046	0.064
	0.072	0.130	0.055	0.075
	0.080	0.114	0.069	0.063
	0.049	0.156	0.049	0.059
	0.052	0.165	0.040	0.063
	0.081	0.143	0.058	0.089
	0.067	0.120	0.053	0.056
	0.055	0.127	0.040	0.063
MEAN	0.063	0.141	0.055	0.066
SD	0.017	0.019	0.014	0.010