Center for Advanced Studies in
Measurement and Assessment

CASMA Research Report

Number 23

Comparison of Three IRT Linking Procedures in the Random Groups Equating Design*

Won-Chan Lee
Jae-Chun Ban[†]

February 2007

[*] The authors are grateful to Robert Brennan for his valuable comments on the paper.

[†] Won-Chan Lee is Research Scientist, Center for Advanced Studies in Measurement and Assessment (CASMA), 210 Lindquist Center, College of Education, University of Iowa, Iowa City, IA 52242 (email: won-chan-lee@uiowa.edu). Jae-Chun Ban is Assistant Professor, Chungnam National University, Korea (email: jban@cnu.ac.kr).

Center for Advanced Studies in
Measurement and Assessment (CASMA)
College of Education
University of Iowa
Iowa City, IA 52242
Tel: 319-335-5439
Web: www.education.uiowa.edu/casma

Contents

List of Tables

List of Figures

Abstract

Various applications of item response theory often require scaling to achieve a common scale for item parameter estimates obtained from different groups. This paper used a simulation to explore the relative performance of three IRT linking procedures in a random groups equating design: concurrent calibration with multiple groups, separate calibration with the Stocking-Lord method, and fixed item parameters with score transformation. The simulation conditions used in this paper included six sampling designs, two levels of sample size, and two levels of the number of items. Four of the six sampling conditions represented situations in which the assumption of randomly equivalent groups was violated to different degrees. In general, all three procedures performed reasonably well across conditions, although some procedures showed relatively more error under some conditions. Some advantages and disadvantages of the three linking procedures are discussed.

1 Introduction

In many item response theory (IRT) models, the latent variable is said to be unidentified up to a linear transformation. If an IRT model fits a set of data, then any linear transformation of the latent trait (θ) scale also fits the set of data, provided that the item parameters also are transformed in an analogous manner. Under the three-parameter logistic IRT model (Lord, 1980), the probability of a correct response for item i at the latent trait value θ is given by:

$$p(\theta|a_i, b_i, c_i) = c_i + \frac{1 - c_i}{1 + \exp[-1.7a_i(\theta - b_i)]}, \tag{1}$$

where a_i, b_i, and c_i are item parameters for item i. If θ is linearly transformed by $\theta^* = S\theta + C$, then the following linear transformation of the item parameters would provide the same fitted probabilities: $a_i^* = a_i/S$, $b_i^* = Sb_i + C$, and $c_i^* = c_i$.

Among several possible equating and linking designs are a common-item nonequivalent groups design and a random groups design. In a common-item nonequivalent groups equating design, two forms are created to have a set of items in common and administered to different groups (Kolen & Brennan, 2004). Then, the common items are used as a link to obtain item parameters of the two forms that are on the same scale. The common-item nonequivalent groups equating design or some variant of it is widely employed in many large-scale testing programs largely because such programs require that only one test form be administered per test administration.

In the random groups equating design, examinees are randomly assigned the test form to be administered. Typically, a spiraling process is used to achieve randomly equivalent groups each taking one of two or more parallel forms. Then, the difference between performance of the groups is attributed to differences in difficulty between/among the forms (Kolen & Brennan, 2004). The goal of the IRT scaling (or linking) in the random groups design is to put the item parameters for the new form (or forms) in the current administration on the ability scale of the group who took the old form in the previous administration. The old form, sometimes called an anchor form, is used to link the new form to the scale of the previous form.

There exist many IRT linking procedures. The two most popular methods are concurrent and separate calibration. As described in the next section, concurrent calibration (with multiple groups) involves estimating parameters using all data simultaneously to obtain a common IRT scale, while the separate calibration procedure involves estimating item parameters separately for each form and then using the linear relationship of the parameter estimates to transform one set of parameter estimates to the scale of the other form. These two general procedures have been compared in many previous studies (Petersen, Cook, & Stocking, 1983; Wingersky, Cook, & Eignor, 1986; Kim & Cohen, 1998; Béguin, Hanson, & Glas, 2000; Béguin & Hanson, 2001; Hanson & Béguin, 2002; Kim & Kolen, 2006).

However, previous research has focused primarily on a common-item nonequivalent groups equating design, and little research has been done comparing the concurrent and separate calibration procedures under the random groups equating design. This paper compares three alternative procedures for obtaining a common scale in a random groups equating design: concurrent calibration with multiple groups (CC), separate calibration with the Stocking-Lord method (SC), and fixed item parameters with score transformation (FP). The FP procedure has not been previously studied.

Using various simulation conditions, the relative performance of the three linking procedures is explored, especially when the assumption of randomly equivalent groups is violated. In addition, the effects of sample size and the number of items on the relative performance of the procedures are examined. All simulated data are calibrated and scored using the computer program BILOG-MG (Zimowski, Muraki, Mislevy, & Bock, 1996). The computer program ST (Hanson & Zeng, 1995) is used for the Stocking-Lord transformation.

2 IRT Linking Procedures

The equating linkage plan considered in this paper is shown in Table 1. Form A administered at Time 1 is the base form from which the basic scale is determined, and item parameters for all other forms are to be put on the scale of Time 1 through an anchor form designated in a parenthesis. In Time 2, the new form, Form B, is administered with Form A in a spiraled administration. Here Form A serves as the anchor form. Likewise, Form B serves as the anchor form at Time 3, and Form C as a new form is administered along with Form B using a spiraling process.

Table 1: A Simple Linkage Plan Using a Random Groups Equating Design

Administration Time	Forms	
Time 1	A	
Time 2	(A)	B
Time 3	(B)	C
Time 4	(C)	D

Focusing on Time 1 and Time 2 only, the purpose of scaling is to put the item parameter estimates for Form B administered at Time 2 on the same scale as Form A administered at Time 1. Within Time 2, the item parameter estimates for Form A and Form B are on the same scale because the two groups taking the two forms are randomly equivalent. The linking design employed in this study mimics this particular linking situation, in which there are three sets of data: Form A administered to Group 1 (A1), Form A administered to Group 2 (A2), and Form B administered to Group $2'$ (B2) (see Figure 1). In this linking design, Group 2 and Group $2'$ do not have any linking items and are assumed to be randomly equivalent (i.e., the two samples are from the same population).

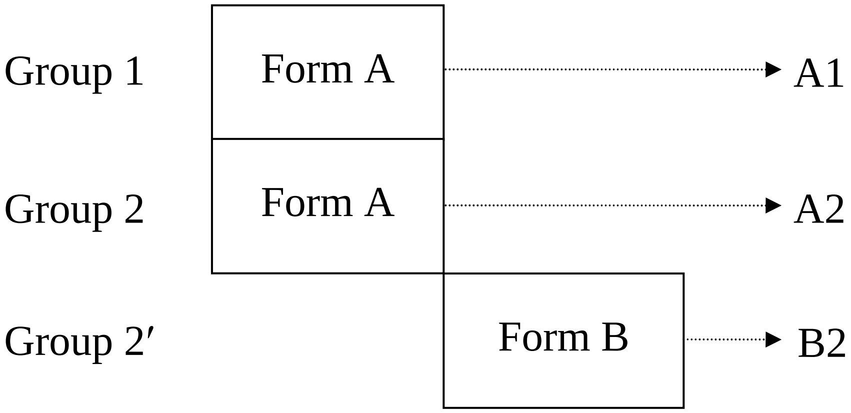

Figure 1: Linking Design Used for Simulations

A2 is used as a link to put the item parameter estimates for B2 on the scale of A1. The three linking procedures employed in this study are briefly discussed next.

2.1 Concurrent Calibration with Multiple Groups (CC)

For CC the item parameters for all items on the three forms (A1, A2, and B2) are estimated simultaneously in a single run of the IRT estimation. This requires estimation software, such as BILOG-MG, that can handle multiple group estimation, in which the means and standard deviations of the latent variables for the groups are allowed to differ. The A1 data serve as the base group in which the mean and standard deviation of the latent variable distribution are fixed at 0 and 1. In the concurrent calibration process, the scale transformation of A2 and B2 is determined based on the comparison of the two estimated theta distributions for A1 and A2. CC and FP (discussed later) have a common property in the sense that both make use of the theta distributions to identify the scale transformation.

2.2 Separate Calibration with the Stocking-Lord Method (SC)

SC uses two sets of item parameter estimates for A1 and A2 obtained from two separate calibrations to estimate the linear scale transformation function. There exist several IRT linking methods for separate calibration as described in Kolen and Brennan (2004): Mean/Mean, Mean/Sigma, Stocking-Lord (Stocking & Lord, 1983), and Haebara (1980). The Mean/Mean and Mean/Sigma methods use moments of the item parameter estimates to obtain the scale transformation parameters. By contrast, the Heabara and Stocking-Lord methods estimate the scale transformation using item characteristic curves or test characteristic curves, respectively. Previous research suggests that the Heabara and Stocking-Lord methods should be preferred over the moment methods (Kolen & Brennan, 2004; Hanson & Béguin, 2002; Kim & Kolen, 2006). In this paper, the Stocking-Lord method is employed.

In the SC condition, the item parameter estimates for A1 and A2 are used to find the scale transformation parameters, S and C, such that the difference between the test characteristic curves for the two sets of item parameter estimates is minimized. The estimated scale transformation parameters, S and C, are then applied to item parameter estimates for B2 to put them on the scale of the item parameters for A1.

2.3 Fixed Item Parameters with Score Transformation (FP)

FP is similar to the linking method called common population linking when applied to the common-item nonequivalent groups equating design (Muraki, Hombo, & Lee, 2000). In essence, FP and common population linking use the estimated latent variables to find the linear transformation function. First, the item parameters for A1 are estimated. Then, A1 item parameter estimates are used to score A2 data constraining the item parameters in the anchor form to be identical. The resultant theta estimates for A2 are on the scale of A1. The next step is to calibrate and score A2 data without imposing any constraint on items. Then, the difference between the two sets of theta estimates for A2 is taken as a direct indication of the difference in group performance between Group 1 and Group 2.

The transformation parameters are estimated using the linear relationship between the two sets of the theta estimates based on the following formulas (Kolen & Brennan, 2004): $S = \sigma(\theta_{Group1})/\sigma(\theta_{Group2})$ and $C = \mu(\theta_{Group1}) - S\mu(\theta_{Group2})$. Under the assumption that A2 and B2 are already on the same scale, the estimated scale transformation parameters, S and C, are then applied to item parameter estimates for B2 to put them on the scale of A1.

As mentioned earlier, FP and CC share a common characteristic of using the theta distributions to determine the scale of measurement, while FP and SC also have something in common in that both procedures are based on the pairwise linking process rather than one simultaneous scaling process. The similarity of the linking procedures may affect their relative performance in a somewhat systematic manner.

3 Method

Real data sets were used to simulate data that are as realistic as possible. The source of data is described first, followed by a discussion of simulation procedures and evaluation criteria.

3.1 Data

This study used data from two 75-item ACT Assessment (ACT, 1997) English forms. Randomly equivalent groups of about 3000 examinees took the two alternate forms of the test, and the item parameters are estimated assuming a three-parameter logistic IRT model. These estimated item parameters for the two forms (Form A and Form B) were used as generating item parameters to simulate data. The generating item parameters for Form A and Form B, respectively, are presented in Tables 2 and 3.

3.2 Simulation

The latent trait variables (θ) were sampled for A1 from a normal distribution with mean 0 and standard deviation 1, which is denoted N(0,1). Four sets of random theta values for A2 were sampled from N(0,1), N(1,1), N(.8,1), and N(.5,1) distributions. Four other sets of random theta values were sampled for B2 from N(0,1), N(1,1), N(.8,1), and N(.5,1) distributions. Thus, a total of nine samples were drawn. Table 4 displays the means and standard deviations of the theta values drawn for the nine samples. The samples of A1 from N(0,1) and samples of A2 and B2 from N(1,1) were used to represent the random groups equating design in which Group 1 and Group 2 (and $2'$) are from different populations, and Group 2 and Group $2'$ are randomly equivalent coming from the same population. This particular condition is denoted M0/1/1 indicating that the three samples, A1, A2, and B2, are in order from normal distributions with means 0, 1, and 1, respectively. There are a total of six combinations of sampling conditions: M0/0/0, M0/1/1, M0/.8/1, M0/1./8, M0/.5/1, and M0/1/.5. The sampling conditions with different means for A2 and B2 represent the case in which the assumption of randomly equivalent groups is violated. The mean values of .8 and .5 are used to differ the degree of violation. For all sampling conditions including those with different means for A2 and B2, the estimation procedures treat A2 and B2 as if they are from the same population to examine the effect of the assumption violation. For the condition of M0/0/0, strictly speaking, scaling is not necessary, but it is included in the simulation design to see whether the scaling makes any difference.

Fifty random item response data sets were generated for each of the nine samples using the same theta values for each sample. Thus, there were 50 triplets of random samples for each of the six combinations of sampling conditions. All three linking procedures were applied to each of the 50 triplets for all six sampling conditions. Two different levels of sample size were considered in this paper: 3000 and 500 per group. The 500 sample size condition used the first 500 of the 3000 examinees per sample. In order to examine the effect of the smaller number of items, the full-length test forms (75 items) were reduced to 25-item forms. For each of the two forms, the generating b parameters were rank ordered, and then 1st, 4th, 7th,..., and 73rd items were selected.[1] Then,

[1] The items in the original test forms are passage-based. The issue of selecting items based on the b parameters without considering passages is mostly irrelevant to this study because

using the same nine samples used to generate random data sets for the full-length forms, 50 triplets of random samples for each of the six combinations of sampling conditions were created for the shorter test forms.

The computer program BILOG-MG was used for item calibration and theta scoring. The prior distribution of theta was assumed to be a standard normal, which is the default option of BILOG-MG. Maximum likelihood (ML) estimation was used for estimating theta parameters. As discussed later, the estimation procedure for the latent parameters had a great impact on the performance of FP. Appendix A provides control files used to obtain item parameter estimates for the CC and SC procedures, and theta estimates for the FP procedure with fixed item parameters. In summary, the factors that were investigated in this paper include: (a) six sampling conditions, (b) three linking procedures, (c) two levels of sample size for each group (n = 3000 and 500), and (d) two levels of the number of items per form (k = 75 and 25). Therefore, there were a total of 72 conditions studied (6 x 3 x 2 x 2). In addition, the FP procedure was replicated over 24 conditions (6 x 2 x 2) using the Baysian expected a posteriori (EAP) method for estimating theta parameters, which was compared to the FP procedure using the ML theta estimates.

3.3 Evaluation Criteria

Each of the 50 sets of item parameter estimates for B2 in each condition should be on the same scale as the generating item parameters, if the B2 item parameters are properly placed on the A1 scale. This is because of the fact that the random data for A1 are generated from N(0,1), and BILOG-MG uses N(0,1) as the default prior for the latent distribution. Since the main purpose of the linking is to put the Form B item parameter estimates on the A1 scale, the Form B items can be evaluated in terms of how close the estimated item parameters are to the generating item parameters.

The evaluation criterion used in this paper was based on the item characteristic curves (ICCs), which was previously employed by Hanson and Béguin (2002). The ICC criterion evaluates how close the estimated item characteristic curves are to the true item characteristic curves for the Form B items. The ICC criterion for item i is given by

$$\begin{aligned}\frac{1}{50}\sum_{r=1}^{50}\int_{-\infty}^{\infty}\Big[p(\theta|a_i,b_i,c_i)-p(\theta|\hat{a}_{ir},\hat{b}_{ir},\hat{c}_{ir})\Big]^2 f(\theta)d\theta = \\ \int_{-\infty}^{\infty}\left[p(\theta|a_i,b_i,c_i)-m_i(\theta)\right]^2 f(\theta)d\theta + \\ \frac{1}{50}\sum_{r=1}^{50}\int_{-\infty}^{\infty}\Big[p(\theta|\hat{a}_{ir},\hat{b}_{ir},\hat{c}_{ir})-m_i(\theta)\Big]^2 f(\theta)d\theta, \qquad (2)\end{aligned}$$

the selected items only serve to provide generating item parameters and the IRT model fit is not of interest here.

where $p(\theta|a_i, b_i, c_i)$ is the ICC for three-parameter logistic IRT model [i.e., Equation 1] using the generating item parameters; $p(\theta|\hat{a}_{ir}, \hat{b}_{ir}, \hat{c}_{ir})$ is the ICC using the estimated item parameters for item i from replication r; $m_i(\theta) = (1/50)\sum_{r=1}^{50} p(\theta|\hat{a}_{ir}, \hat{b}_{ir}, \hat{c}_{ir})$; and $f(\theta)$ is the density for θ. The term on the left hand side of Equation 2 is the mean squared error (MSE) of the estimated ICC for item i, which can be decomposed into the squared bias (SB) and variance (VAR) as shown on the right side of the equation. Average values of these three indices over 75 (or 25 for shorter test form conditions) items were computed for each condition.

The integrals in Equation 2 were evaluated using a Monte Carlo integration algorithm: $I = \int_a^b g(x)dx \simeq [(b-a)/K]\sum_{k=1}^{K} g(x_k)$, where $x_k, k = 1, \ldots, K$ is a set of K (=1000) random deviates drawn from a uniform distribution with an interval of $a = -5$ and $b = 5$. As for the Hanson and Béguin (2002) study, two weight functions, $f(\theta)$, were initially used to compute the three indices: (a) a unit weight, $f(\theta) = 1$, for each Monte Carlo uniform random deviate, and (b) a weight associated with the N(0,1) density function. Since the results based on the two weight functions were very similar, only the results using the normal-density weights are reported here. The results are reported in terms of the total error (MSE), SB, and VAR for each condition.

4 Results

For all BILOG-MG runs, the maximum number of EM cycles of 30 was used to insure convergence (see Appendix A), and all runs converged. The average values for MSE, SB, and VAR for the condition of full-length test forms ($k = 75$) are summarized in Table 5. The column labeled "NO" indicates separate calibration without any scaling. The NO condition can serve as the baseline criterion.

A few general observations can be made from Table 5. First, the MSE for the conditions with $n = 3000$ tends to be lower than the MSE for the conditions with $n = 500$ primarily due to the lower VAR. Second, the MSE gets larger as the degree of the assumption violation gets more severe. Third, the performance of SC and FP tends to be extremely similar while CC shows a somewhat distinctive pattern. Across all conditions, the VAR values for the three linking procedures are congruent. Most differences are found in the SB, which results in differences in the MSE. A more detailed discussion by sampling conditions is provided next.

When all three samples for A1, A2, and B2 are from the same population (i.e., M0/0/0 condition), scaling does not seem to make any improvement. Note that, in their simulation study, Hanson and Béguin (2002) reported that some scaling procedures produced less error than no scaling under the common-item nonequivalent groups equating design even when the two groups are equivalent. As expected, no scaling produces large bias when the two populations for Group 1 and Group 2 (and 2$'$) differ (i.e., all the other conditions but M0/0/0), except for the M0/1/.5 condition with the sample size of 500 where the assumption of randomly equivalent groups is severely violated.

For the M0/1/1 condition, in which Group 2 and Group 2$'$ are randomly equivalent, the results for the three procedures do not seem to differ substantially. Especially, for the small sample size condition, the performance of the three procedures is very similar. When $n = 3000$, however, SC and FP tend to slightly outperform CC in terms of the SB and MSE.

When the assumption of randomly equivalent groups is violated, each linking procedure tends to show a particular pattern. In general, SC and FP exhibit a clear pattern of outperforming CC regardless of the sample size when the Group 2$'$ (B2) mean is higher than the Group 2 (A2) mean (i.e., M0/.8./1 and M0/.5/1 conditions). This is primarily due to the SB. By contrast, CC clearly outperforms SC and FP when the Group 2 mean is higher than the Group 2$'$ mean (i.e., M0/1/.8 and M0/1/.5 conditions). The direction of the mean difference between Group 2 and Group 2$'$ appears to have a great impact on the amount of error. A comparison of M0/.8/1 versus M0/1/.8 and M0/.5/1 versus M0/1/.5 reveals that more error is produced when the Group 2$'$ mean is larger than the Group 2 mean (i.e., M0/.8/1 and M0/.5/1 conditions) even though the absolute amount of the mean difference is the same. This is largely because of the fact that the variance of b-parameter estimates tends to be larger when estimated for a more able group. Consider a hypothetical example presented in Table 6. In this example, only the b parameters are considered and linking is conducted using the SC procedure with the mean/sigma method (see Kolen & Brennan, 2004). Under the assumption that the variance of b-parameter estimates is larger for more able groups, the two conditions, M0/.5/1 and M0/1/.5, are compared in terms of the transformed mean b-parameter estimate for B2 using the transformation parameters, S and C, obtained from linking A2 to A1. As shown in the table, the mean b under the M0/.5/1 condition is transformed to .4878, which is larger than the result for the M0/1/.5 condition (i.e., -.4762) in the absolute sense. These two numbers are indeed the bias due to ignorance of the fact that the two groups are from different populations. This somewhat oversimplified example demonstrates that the direction of the mean difference between Group 2 and Group 2$'$ treating them as if they are from a single population can affect the performance of the linking procedures primarily due to different, but consistent variance of b-parameter estimates for different ability groups.

One exception is the condition of M0/1/.5 with n=500, which shows the largest error among all conditions. In this case, no-scaling works better than the scaling procedures. Note that, for the M0/1/.5 condition, it is expected that no-scaling and the scaling procedures will perform similarly. Roughly speaking, the scaling of A2 to A1 under the M0/1/.5 condition adjusts the mean difference of 1.0. Then, applying the same scaling parameters to B2, which has a mean of .5, will in effect lead to an over-adjustment by .5. No-scaling of B2, of course, will retain the difference of .5. Nonetheless, the relatively better performance of no-scaling for the condition of M0/1/.5 with n=500 appears to be because of the sample mean of B2 (=.412, see Table 4) that is substantially lower than the population mean of .5, which causes an over-adjustment for the scaling procedures by approximately .6, and no-scaling retains the difference of .4.

The results based on shorter test forms ($k = 25$) are displayed in Table 7. Almost the same general observations can be made for the $k = 25$ conditions as for the $k = 75$ conditions. Some unique observations, however, can also be made from Table 7. The tendency of the relatively larger MSE and SB for CC when $n = 3000$ is not found any more. In general, the MSE and SB for the three linking procedures in the $k = 25$ conditions are more similar to each other than they are in the $k = 75$ conditions. There exists a clear pattern that SC performs better than FP when the mean of Group $2'$ is greater than that of Group 2, and on the contrary, FP performs better than SC when the mean of Group 2 is greater than that of Group $2'$.

Table 8 presents the results for the FP procedure when the ML and EAP theta estimates are used for linking. In the M0/1/1 condition, the ML estimates tend to produce lower SB and MSE, especially for the shorter test forms. Overall, the ML method shows lower MSE for the conditions of M0/1/1, M0/.8/1, and M0/.5/1 mainly because of the lower SB. By contrast, the EAP estimates show less MSE for the M0/1/.8 and M0/1/.5 conditions. This appears to be highly related to the EAP estimates being shrunk in variance, which is a generic property of EAP estimation (Mislevy & Bock, 1990). The shrinkage in the variance of EAP estimates leads to potential bias at both tails of the theta distribution. The relatively poorer performance of the EAP method under the M0/1/1, M0/.8/1, and M0/.5/1 conditions is due to the fact that the EAP estimates obtained by fixing the item parameter estimates from A1 and scoring the A2 data are underestimated because the estimated theta distribution shifts upward and shrinkage to the mean occurs more at the higher end of the theta distribution. For example, for the M0/.8/1 condition, the mean theta value for A1 would be close to 0, and scoring A2 with item parameters fixed at the values obtained from calibration of A1 would be expected to yield a mean theta value close to .8. Yet, the mean EAP estimates will be smaller than .8.[2] This results in a less severe transformation function applied to the B2 data than it is supposed to be. The same logic can be applied to explain the opposite phenomenon, that is, the relatively better performance of the EAP method for the M0/1/.8 and M0/1/.5 conditions.

5 Discussion

Various applications of IRT, such as test equating, DIF analysis, and computer adaptive testing, often require IRT scaling to achieve a common scale for item parameter estimates obtained from different people and at different times. Although there exist many IRT scaling methods and their variations, little research has provided comprehensive comparison of the various methods. Moreover, previous research has not shown consistent conclusions about which method to prefer. This paper should be considered as an additional effort to help measurement practitioners evaluating linking procedures suitable for their

[2] The ML estimates also show some shrinkage to the mean because of the ceiling effect, but the degree of shrinkage is less than that of the EAP estimates.

practical situations. In particular, this paper compared three IRT linking procedures (two of them have been widely known and the other one is relatively unknown) under the random groups equating design, which has not been investigated in the previous studies despite its popularity in many real testing programs. Most previous studies examined the common-item nonequivalent groups equating design.

This paper employed simulations to explore the relative performance of the CC, SC, and FP procedures using several simulation conditions including six sampling designs, two levels of sample size, and two levels of the number of items. The six sampling conditions involved two designs in which the assumption of randomly equivalent groups was met, and four other conditions in which the assumption was violated to different degrees. The performance of CC relative to that of SC and FP is summarized as follows: 1) CC performed worse than the other two procedures when the sample size and number of items are large and the randomly equivalent groups assumption is met (i.e., M0/1/1 condition); 2) CC produced relatively large error when the randomly equivalent groups assumption was violated such that the mean of Group $2'$ was higher than that of Group 2 (i.e., M0/.8/1 and M0/.5/1 conditions); and 3) CC showed less error when the assumption was violated such that the mean of Group $2'$ was lower than that of Group 2 (i.e., M0/1/.8 and M0/1/.5 conditions).

The SC and FP procedures performed more similarly to each other than to CC, largely because of their characteristic of separate and multiple estimations. Their behaviors were slightly different, however, when the number of items in the forms was small. The relative performance of these two procedures for the shorter forms interacted with the sampling conditions. The differences were not very substantial though. The difference between the performance of CC and the other two procedures (SC and FP) might also be attributable to the different sources of potential bias. The CC estimation process involves data from A2 and B2 that do not have any linking items or examinees, although they are assumed to be from the same population. This nature of a disconnectedness between the data from A2 and B2 in a random groups equating design for concurrent calibration might be a source of potential bias, either positive or negative depending upon the samples. The SC and FP procedures also have some potential bias due to multiple pairwise linking (Muraki et al., 2000). Different sources of bias may well affect the performance of the linking procedures.

One distinct advantage of CC is that it requires only one estimation process. On the other hand, all response data must be available at the time of calibration and linking, which is not always feasible in practice. In some cases, only item parameter estimates are available and linking should be conducted without any access to the old form response data. In that case, the SC or FP would be more viable options.

For both the FP and SC procedures, the items in the anchor form are constrained to be identical. This would work fine as long as the item parameters are believed not to drift across time. When the items are discovered to exhibit some parameter drift, however, the CC procedure could be used incorporating item drift parameters in estimation (Muraki et al, 2000). Doing so requires estimation software such as BILOG-MG and PARSCALE (Muraki & Bock, 1998). In the comparison of the SC and FP procedures, common population linking is sometimes claimed to be preferred to common item linking procedures because the latent distributions are less sensitive to changes in individual items (Donoghue & Mazzeo, 1992; Muraki et al., 2000). Thus, when item parameter drift is suspected, FP might be preferred to SC.

In the comparison of ML and EAP theta estimates for the FP procedure, it was found that the ML estimates produced less error when the assumption of randomly equivalent groups was met (i.e., M0/1/1 condition). When the assumption was violated, the performance of the ML and EAP estimates depended on the direction of the mean difference between Group 2 and Group $2'$. In general, the ML method would be preferred because of its better performance under a more "typical" and important equating situation (i.e., the assumption is met), and its tremendous comparability with the SC procedure, which is known to be very accurate and most widely used in practice. Moreover, the better performance of the EAP method in some conditions was primarily due to the characteristic of shrinkage rather than "accuracy" of estimation per se.

Since the relative performance of different linking procedures varies according to the measurement conditions, it does not seem to be sensible to draw a conclusion in favor of any one approach in all occasions. General discussions presented thus far, however, might serve as practical guidelines in choosing a procedure in various linking situations. Also, as Hanson and Bguin (2002) suggested, it would be beneficial, in practice, to apply multiple linking procedures and compare the results. Doing so would help understand various issues and aspects of the linking situation in hand and lead to a choice of the most appropriate linking method. Drawing conclusions from the present simulation study should be limited due to the small number of conditions investigated in this paper. For example, in this paper, the data were generated based on the model that was also used in the estimation. The misfit of the model to the data may affect the relative performance of the linking procedures. A further simulation study might involve generating data from distributions other than the normal such as a skewed distribution, or multidimensional IRT models could be used to impose some degree of model misfit on the data. One other limitation of the present study was the use of only one computer software package BILOG-MG. A useful further research project would involve comparison of different estimation software with options available for multiple group estimation, such as MULTILOG (Thissen, 1991) and ICL (Hanson, 2002).

Table 2: Generating Item Parameters for Form A Used for Simulations

Item	*a*	*b*	*c*	Item	*a*	*b*	*c*
1	.325	-1.396	.285	39	.051	.254	.271
2	.826	.172	.305	40	.793	.592	.209
3	.595	-1.479	.256	41	.980	1.568	.135
4	.656	-2.065	.291	42	.794	.495	.178
5	.874	-.190	.403	43	.912	1.225	.372
6	.569	.099	.441	44	.891	-1.034	.309
7	.500	1.732	.130	45	.629	-.264	.218
8	.844	-.360	.282	46	.042	-.445	.151
9	.770	-.190	.254	47	.624	1.656	.143
10	.559	.330	.330	48	.599	.116	.145
11	.792	.521	.404	49	.395	.519	.142
12	.877	-.819	.366	50	.765	-.303	.192
13	.611	.133	.358	51	.199	-.113	.219
14	.511	.640	.240	52	.712	-1.280	.321
15	.739	.226	.346	53	.552	-.668	.094
16	.938	1.430	.411	54	.824	-.377	.143
17	.705	-.033	.458	55	.840	-.322	.097
18	.785	-1.331	.296	56	.144	-.581	.271
19	.824	-.851	.239	57	.490	.200	.312
20	.827	.918	.470	58	.282	.259	.248
21	.692	-.326	.178	59	.535	.685	.328
22	.647	-.437	.211	60	.201	-.178	.341
23	.935	.593	.056	61	.472	.197	.145
24	.845	-.984	.272	62	.046	-.185	.106
25	.818	.878	.322	63	.745	-.789	.248
26	.365	-.684	.355	64	.774	-.662	.186
27	.605	-.989	.151	65	.842	1.619	.209
28	.736	-.985	.220	66	.976	-.888	.268
29	.735	-.067	.311	67	.968	-.546	.256
30	.865	.676	.422	68	.440	-.530	.190
31	.674	-1.110	.282	69	.950	-.328	.250
32	.563	.268	.173	70	.303	1.360	.147
33	.650	-.757	.272	71	.811	.529	.160
34	.759	-1.206	.263	72	.959	-.105	.192
35	.439	-.365	.486	73	.686	.631	.153
36	.832	-.669	.235	74	.903	-.274	.225
37	.308	-2.225	.249	75	.646	.631	.087
38	.602	-1.035	.140				

Table 3: Generating Item Parameters for Form B Used for Simulations

Item	*a*	*b*	*c*	Item	*a*	*b*	*c*
1	.400	-1.300	.334	39	.701	-.627	.215
2	.335	-3.900	.249	40	.721	-1.634	.318
3	.401	-.030	.254	41	.901	-.060	.117
4	.783	.685	.193	42	.630	-1.164	.324
5	.376	-3.345	.255	43	.680	-.600	.107
6	.485	-1.066	.214	44	.536	.619	.139
7	.618	-1.577	.405	45	.419	-.654	.164
8	.849	-.496	.254	46	.642	-.994	.174
9	.929	-1.589	.285	47	.746	-.663	.301
10	.770	.187	.121	48	.758	-.455	.102
11	.304	-1.691	.243	49	.740	-.518	.301
12	.521	.376	.226	50	.643	-.131	.152
13	.433	-1.688	.302	51	.711	-.314	.213
14	.613	-.164	.170	52	.482	.354	.260
15	.227	-.229	.399	53	.584	-1.108	.219
16	.513	-2.650	.236	54	.723	-1.027	.280
17	.490	-.843	.392	55	.930	.074	.084
18	.773	1.764	.192	56	.709	.533	.354
19	.933	-.207	.360	57	.686	.318	.137
20	.029	.118	.317	58	.242	-.013	.177
21	.794	-.407	.130	59	.822	-.058	.417
22	.479	-1.554	.316	60	.972	-.642	.234
23	.469	-1.954	.216	61	.786	1.273	.352
24	.740	-.893	.335	62	.964	-.123	.268
25	.417	-1.926	.289	63	.766	-.003	.255
26	.003	-.483	.254	64	.760	.983	.193
27	.747	-.984	.255	65	.865	-.330	.290
28	.599	.438	.187	66	.132	-.259	.272
29	.558	-.802	.176	67	.827	.859	.129
30	.622	-1.431	.311	68	.922	.159	.249
31	.905	-1.834	.183	69	.979	-.138	.317
32	.673	.808	.087	70	.021	.459	.142
33	.485	-1.250	.213	71	.958	-.066	.459
34	.678	-.236	.189	72	.416	.572	.258
35	.447	.206	.221	73	.385	.050	.366
36	.630	-.811	.196	74	.102	.166	.415
37	.721	-.097	.210	75	.880	.495	.098
38	.663	-2.150	.241				

Table 4: First Two Moments of Theta Values for the Nine Samples Used for Simulations

	Sampling	Mean	SD
$n = 3000$			
	A1: N(0,1)	.021	.975
	A2: N(0,1)	.003	1.013
	A2: N(1,1)	.986	1.019
	A2: N(.8,1)	.785	.990
	A2: N(.5,1)	.494	.998
	B2: N(0,1)	.024	.994
	B2: N(1,1)	.972	.994
	B2: N(.8,1)	.800	.993
	B2: N(.5,1)	.508	.991
$n = 500$			
	A1: N(0,1)	.003	1.013
	A2: N(0,1)	-.013	.981
	A2: N(1,1)	1.005	1.022
	A2: N(.8,1)	.784	.997
	A2: N(.5,1)	.453	.928
	B2: N(0,1)	.036	.991
	B2: N(1,1)	.960	.991
	B2: N(.8,1)	.861	.994
	B2: N(.5,1)	.412	.972

Table 5: Average MSE, Squared Bias, and Variance Using Full Length ($k = 75$) Test Forms (x 1,000)

		$n = 3000$				$n = 500$			
Sampling		CC	SC	FP	NO	CC	SC	FP	NO
M0/0/0									
	MSE	.002	.002	.002	.002	.009	.010	.010	.009
	SB	.000	.001	.001	.000	.002	.002	.002	.002
	VAR	.002	.002	.002	.002	.007	.007	.008	.007
M0/1/1									
	MSE	.006	.004	.004	.261	.015	.015	.016	.260
	SB	.004	.002	.001	.260	.004	.004	.005	.254
	VAR	.003	.003	.003	.001	.011	.011	.012	.005
M0/.8/1									
	MSE	.022	.015	.015	.261	.029	.022	.022	.260
	SB	.020	.013	.012	.260	.019	.012	.012	.254
	VAR	.002	.002	.002	.001	.010	.010	.010	.005
M0/1/.8									
	MSE	.007	.011	.011	.184	.017	.020	.021	.214
	SB	.005	.008	.008	.183	.005	.009	.009	.208
	VAR	.003	.003	.003	.001	.012	.011	.012	.006
M0/.5/1									
	MSE	.089	.077	.077	.261	.089	.073	.074	.260
	SB	.088	.076	.076	.260	.081	.065	.066	.254
	VAR	.002	.002	.002	.001	.008	.008	.008	.005
M0/1/.5									
	MSE	.057	.067	.067	.077	.102	.114	.114	.058
	SB	.054	.065	.065	.076	.091	.103	.103	.052
	VAR	.003	.003	.003	.001	.011	.011	.011	.006

Table 6: A Hypothetical Example of Linking with Different Directions of Mean Difference When the Assumption is Violated

	M0/.5/1 Condition		
	A1: N(0,1)	A2: N(.5,1)	B2: N(1,1)
Mean(b)	0.00	0.50	1.00
SD(b)	0.80	0.82	0.84
$S = \sigma(b_{Group1})/\sigma(b_{Group2}) = .8/.82 = .9756$			
$C = \mu(b_{Group1}) - S\mu(b_{Group2}) = 0 - .9756(.5) = -.4878$			
Applied to B2 at mean(b): $b^* = Sb + C = (.9756)(1.0) - .4878 = .4878$			

	M0/1/.5 Condition		
	A1: N(0,1)	A2: N(1,1)	B2: N(.5,1)
Mean(b)	0.00	1.00	0.50
SD(b)	0.80	0.84	0.82
$S = \sigma(b_{Group1})/\sigma(b_{Group2}) = .8/.84 = .9524$			
$C = \mu(b_{Group1}) - S\mu(b_{Group2}) = 0 - .9524(1.0) = -.9524$			
Applied to B2 at mean(b): $b^* = Sb + C = (.9524)(0.5) - .9524 = -.4762$			

Table 7: Average MSE, Squared Bias, and Variance Using Shorter ($k = 25$) Test Forms (x 1,000)

		$n = 3000$				$n = 500$			
Sampling		CC	SC	FP	NO	CC	SC	FP	NO
M0/0/0									
	MSE	.002	.003	.003	.002	.010	.011	.011	.009
	SB	.000	.001	.001	.000	.002	.003	.003	.002
	VAR	.002	.002	.002	.002	.008	.008	.008	.008
M0/1/1									
	MSE	.006	.005	.006	.255	.017	.018	.017	.251
	SB	.003	.002	.003	.254	.004	.004	.004	.244
	VAR	.003	.003	.003	.001	.014	.013	.013	.006
M0/.8/1									
	MSE	.018	.016	.017	.255	.024	.023	.028	.251
	SB	.015	.013	.014	.254	.012	.011	.010	.244
	VAR	.003	.003	.003	.001	.012	.012	.018	.006
M0/1/.8									
	MSE	.011	.012	.011	.181	.020	.022	.020	.210
	SB	.008	.009	.007	.179	.007	.009	.007	.203
	VAR	.003	.003	.003	.001	.013	.013	.013	.006
M0/.5/1									
	MSE	.081	.076	.079	.255	.076	.069	.074	.251
	SB	.079	.074	.077	.254	.065	.059	.064	.244
	VAR	.002	.002	.002	.001	.010	.010	.010	.006
M0/1/.5									
	MSE	.066	.068	.064	.076	.115	.118	.111	.057
	SB	.062	.065	.060	.074	.103	.106	.099	.050
	VAR	.003	.003	.004	.002	.012	.012	.012	.007

Table 8: Comparison of ML and EAP Theta Estimates for the FP Procedure (x 1,000)

		Full Length Forms ($k = 75$)				Shorter Forms ($k = 25$)			
		$n = 3000$		$n = 500$		$n = 3000$		$n = 500$	
Sampling		ML	EAP	ML	EAP	ML	EAP	ML	EAP
M0/0/0									
	MSE	.002	.002	.010	.010	.003	.002	.006	.010
	SB	.001	.001	.002	.002	.001	.001	.001	.002
	VAR	.002	.002	.008	.007	.002	.002	.005	.008
M0/1/1									
	MSE	.004	.005	.016	.016	.006	.014	.012	.021
	SB	.001	.003	.005	.005	.003	.011	.003	.009
	VAR	.003	.003	.012	.011	.003	.003	.008	.012
M0/.8/1									
	MSE	.015	.022	.022	.028	.017	.037	.021	.040
	SB	.012	.020	.012	.019	.014	.034	.014	.029
	VAR	.002	.002	.010	.010	.003	.002	.007	.011
M0/1/.8									
	MSE	.011	.006	.021	.017	.011	.004	.022	.016
	SB	.008	.004	.009	.006	.007	.001	.014	.005
	VAR	.003	.003	.012	.011	.003	.003	.008	.011
M0/.5/1									
	MSE	.077	.087	.074	.086	.079	.102	.070	.096
	SB	.076	.085	.066	.078	.077	.100	.064	.086
	VAR	.002	.002	.008	.008	.002	.002	.006	.010
M0/1/.5									
	MSE	.067	.051	.114	.094	.064	.029	.106	.067
	SB	.065	.048	.103	.083	.060	.026	.099	.056
	VAR	.003	.003	.011	.011	.004	.003	.008	.011

6 References

ACT, Inc. (1997). *ACT Assessment technical manual.* Iowa City, IA: Author.

Béguin, A. A., Hanson, B. A., & Glas, C. A. W. (2000, April). *Effect of multidimensionality on separate and concurrent estimation in IRT equating.* Paper presented at the Annual Meeting of the National Council of Measurement in Education, New Orleans.

Béguin, A. A., & Hanson, B. A. (2001, April). *Effect of noncompensatory multidimensionality on separate and concurrent estimation in IRT observed score equating.* Paper presented at the Annual Meeting of the National Council of Measurement in Education, Seattle.

Donoghue, J. R., & Mazzeo, J. (1992, April). *Comparing IRT-based equating procedures for trend measurement in a complex test design.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco.

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

Hanson, B. A. (2002). *IRT Command Language (ICL)* [Computer program]. (Available at http://www.b-a-h.com/software/irt/icl/index.html).

Hanson, B. A. & Zeng, L. (1995). *ST: A Computer Program for IRT Scale Transformation.*

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., & Kolen, M. J. (2006). Robustness to format effects of IRT linking methods for mixed-format tests. *Applied Measurement in Education*, *19*, 357–381.

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

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

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

Mislevy, R. J., & Bock, R. D. (1990). *BILOG3: Item analysis and test scoring with binary logistic models* [Computer program]. Mooresville IN: Scientific Software.

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

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

Pertersen, N. S., Cook, L. L., & Stocking, M. L. (1983). IRT versus conventional equating methods: A comparative study of scale stability. *Journal of Educational Statistics*, *8*, 137–156.

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

Thissen, D. (1991). *Multilog user's guide: Multiple, categorical item analysis and test scoring using item response theory* [Computer program]. Chicago: Scientific Software International.

Wingersky, M. S., Cook, L. L., & Eignor, D. R. (1986). *Specifying the characteristics of linking items used for item response theory item calibration.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Zimowski, M. F., Muraki, E., Mislevy, R. J., & Bock, R. D. (1996). *BILOG-MG: Multiple-group IRT Analysis and Test Maintenance for Binary Items* [Computer program]. Chicago: Scientific Software International.

A BILOG-MG Control Card Files

The BILOG-MG control card files used for the three scaling procedures are provided in this appendix. The examples presented in this appendix are based on 3000 examinees' 0/1 item responses on 75 items per group. Concurrent calibration uses all three data sets together (A1, A2, and B2) with A1 being specified as the reference group. The IFN card in estimating latent variables by fixing item parameters for the FP procedure should include the name of the item parameter file, which is a direct output from the BILOG-MG SAVE and PARM options.

`Separate Calibration`

```
    >COMMENT separate calibration
    >GLOBAL NPARM=3,DFN='datafile',NTEST=1,SAVE;
    >SAVE   PARM='parameterfile';
    >LENGTH NITEMS=75;
    >INPUT  NTOT=75,SAMPLE=3000,NALT=4,NID=4;
    >ITEMS  ;
    >TEST   INUM=(1(1)75);
            (4A1,T9,75A1)
    >CALIB  NQPT=40,CYCLE=30;
```

`Concurrent Calibration with Two Groups`

```
    >COMMENT concurrent calibration with two groups
    >GLOBAL NPARM=3,DFN='datafile',NTEST=1,SAVE;
    >SAVE   PARM='parameterfile';
    >LENGTH NITEMS=150;
    >INPUT  NTOT=150,SAMPLE=9000,NALT=4,NID=4,NFORM=2,NGROUP=2;
    >ITEMS  ;
    >TEST   INUM=(1(1)150);
    >FORM1  LENGTH=75,INUMBERS=(1(1)75);
    >FORM2  LENGTH=75,INUMBERS=(76(1)150);
    >GROUP1 GNANE='group1',LENGTH=75,INUMBERS=(1(1)75);
    >GROUP2 GNANE='group2',LENGTH=150,INUMBERS=(1(1)150);
            (4A1,T6,2I1,T9,75A1)
    >CALIB  NQPT=40,CYCLE=30,REF=1;
```

`Fixed Item Parameters and Scoring`

```
    >COMMENT fixed item parameters and ML scoring
    >GLOBAL NPARM=3,DFN='datafile',IFN='fixedparfile',NTEST=1,SAVE;
    >SAVE   SCORE='scorefile';
    >LENGTH NITEMS=75;
    >INPUT  NTOT=75,SAMPLE=3000,NALT=4,NID=4;
```

```
>ITEMS  ;
>TEST   INUM=(1(1)75);
        (4A1,T9,75A1)
>CALIB  ;
>SCORE  METHOD=1;
```