Center for Advanced Studies in Measurement and Assessment

CASMA Research Report

Number 49

Multiple Group IRT Fixed-Parameter Estimation for Maintaining an Established Ability Scale

Seonghoon Kim
Michael J. Kolen[†]

August 2016

[†]Seonghoon Kim is Associate Professor, College of Education, Hanyang University, the Republic of Korea (email: seonghoonkim@hanyang.ac.kr). Michael J. Kolen is Professor, College of Education, University of Iowa, Iowa City, IA 52242 (email: michael-kolen@uiowa.edu).

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

Under the context of a common-item linking design involving multiple examinee groups and multiple test forms, this paper presents a multiple-group (MG) IRT fixed-parameter estimation (FPE) method, which is based on the EM algorithm for finite mixtures. The FPE method that is presented is an extension of the single-group FPE method to MG test data. In a successful MG FPE run, all the parameters of freed items across test forms and the probabilities of the discrete ability distributions for multiple groups are to be properly estimated on the established scale of the fixed items. A simulation study shows that the method that is presented produces more accurate parameter estimates than other estimation and linking methods for estimating freed item parameters and ability distributions on the established scale.

1 Introduction

It is well known that the scale used for measuring ability (denoted as θ) in item response theory (IRT) is determined up to a linear transformation as long as the item parameters are transformed accordingly (de Ayala, 2009; Hambleton, 1989; Lord, 1980; Yen & Fitzpatrick, 2006). In practical applications of IRT, such scale indeterminacy is often solved by "standardizing" the underlying ability distribution for the group of examinees being analyzed (Mislevy & Bock, 1990). The established "0-1" ability scale must be maintained to identify and utilize the invariance property of item parameters across groups. Maintaining the established scale is critical, particularly, for IRT-based test equating, vertical scaling, and development of an item pool (Kim, Harris, & Kolen, 2010; Kolen & Brennan, 2014; Vale, 1986; Young, 2006). When a new test form is developed for test equating, the underlying ability distribution of an examinee group taking the test form and its item parameters need to be put on the established scale. If the new test form contains "old" items and their IRT parameters are well-established, parameters for all other "new" items should be estimated using the scale of the old items.

For such estimation, Kim (2006) presented an effective fixed-parameter estimation (FPE) method that uses the expectation-maximization (EM) algorithm to iteratively estimate both the probabilities of the ability distribution and the parameters of the new items with the old items' parameters being fixed at those values on the established scale (see also Paek and Young, 2005, for FPE). The effectiveness of the FPE method has been supported in many subsequent studies (e.g., Baldwin, Nering, & Baldwin, 2007; DeMars & Jurich, 2012; Keller & Hambleton, 2013; Keller & Keller, 2011). However, Kim's (2006) iterative FPE method has a limitation in that it can be applied only to single-group (SG) test data.

The purpose of this paper is twofold: (a) to extend Kim's (2006) FPE method to multiple-group (MG) test data, which are typically obtained under the common-item nonequivalent groups (CING) equating/linking design (Kolen & Brennan, 2014), and (b) to examine its performance, relative to other estimation and linking methods, as to how accurately it recovers both the item parameters of new test forms and the ability distributions of groups taking the forms on the established scale. The extension of IRT FPE to MG test data uses the general EM algorithm for finite mixtures (Dempster, Laird, & Rubin, 1977; McLachlan & Peel, 2000; Titterington, Smith, & Makov, 1985), assuming the latent ability variable is discrete. In a single successful MG FPE run, all the parameters of freed items across test forms and the probabilities of the discrete ability distributions for multiple groups are to be properly estimated on the established scale of some items with parameters that are fixed.

2 The Need for MG IRT FPE

The need for the MG IRT FPE method may be shown using an exemplar CING equating situation in which new ability scales from MG test data are linked to the established scale. Suppose that two test forms, Form 1 and Form 2, are administered to two nonequivalent examinee groups, Group 1 and Group 2, respectively (Figure 1 in a later section provides a picture of this design). Form 1 consists of both "old" items and "new" items. Form 2 consists of all "new" items. The old items of Form 1 have come from an item pool, which resulted from more than one previous test sessions, so their parameter estimates are known and placed on the established ability scale of the item pool. The old items play the role of "linking" items between the current MG test data and the item pool. In addition, the two forms have some new items in common. Once the MG test data are obtained, the task of interest is to estimate the new item parameters of the two forms and the ability distributions of the two groups so that all the estimates may be put on the established scale. The task can be accomplished by at least three different approaches. To explain the approaches, denote the established scale, the 0-1 scale from Group 1 taking Form 1, and the 0-1 scale from Group 2 taking Form 2 as θ_0, θ_1, and θ_2, respectively.

The first approach for item parameter estimation is MG FPE for the combined data of Forms 1 and 2, with the parameters of the old linking items being fixed at their values on the θ_0 scale. Unlike the other two approaches, the MG FPE approach requires a "single" IRT estimation computer run to place all parameter estimates from the MG test data on the θ_0 scale.

The second approach is MG 0-1 estimation and linking, involving two steps: (a) to conduct MG estimation (also known as "concurrent estimation") with the combined data of Forms 1 and 2 and with Group 1 being designated as the base/reference group, and (b) to use the old linking items to find the linking coefficients A_1 (slope) and B_1 (intercept) of a linear relation $\theta_0 = A_1\theta_1 + B_1$ so that all parameter estimates on the 0-1 scale from MG estimation are transformed onto the θ_0 scale using this linear relation. The MG 0-1 estimation in the first step has been presented by Bock and Zimowski (1997) and is described in this paper using the EM algorithm for finite mixtures. In the second step, the linking coefficients can be found by using moment methods (Loyd & Hoover, 1980; Marco, 1977), characteristic curve methods (Haebara, 1980; Stocking & Lord, 1983), or other methods (Divgi, 1985; Ogasawara, 2001). It has been shown that the characteristic curve methods typically produce more accurate linking results than other linking methods in developing a common ability scale (Kolen & Brennan, 2014).

The third approach is SG separate 0-1 estimation and linking (simply referred to as "SG 0-1 estimation and linking"), which involves three steps: (a) conduct SG separate 0-1 estimation for each form, (b) use the old linking items or the common items between forms to find the A_1 and B_1 of a "θ_1 to θ_0" linear transformation and the A_2 and B_2 of a "θ_2 to θ_1" linear transformation, and (c) chain, from θ_2 to θ_0, the linear transformations. In the third step, the item parameter estimates and ability points from Form 1 and Form 2 are

transformed onto the θ_0 scale using the linear relations $\theta_0 = A_1\theta_1 + B_1$ and $\theta_0 = (A_1A_2)\theta_2 + (B_1 + A_1B_2)$, respectively.

The need for the MG FPE approach can be justified by its possible advantages over the other two approaches. First, compared to the second approach that involves MG 0-1 estimation and scale linking, the MG FPE approach has two benefits. One is pertinent to the common items, for which only one set of parameter estimates is obtained for Forms 1 and 2 and, because of a doubled sample size, the estimation errors are relatively smaller than those for the non-common new items (see Hanson & Béguin, 2002; Kim & Kolen, 2006). The other is relevant to all items of Forms 1 and 2, in that item parameter estimates do not contain estimation error due to linking a new 0-1 scale to the established scale.

Second, the MG FPE approach can avoid several disadvantages of the third "SG 0-1 estimation and linking" approach. The third approach requires many computer runs for separate 0-1 estimation and scale linking, it results in two sets of parameter estimates for the common items, it yields relatively larger estimation errors for the common items, and it makes the item parameter estimates of the form positioned later in linkage (Form 2 in this exemplar situation) absorb the accumulated errors induced by chaining scale transformations.

In sum, the MG FPE approach presented in this paper is expected to be efficient and effective for achieving the purpose of analyzing MG test data while maintaining the established ability scale. The MG FPE approach needs a single computer run with the MG test data. It does not involve scale transformations and thus can avoid estimation errors induced by those processes. It yields only one set of parameter estimates for the common items between test forms. In addition, as far as the common items are concerned, the MG FPE method uses all examinees responding to the items and so estimates their parameters using full information, likely leading to less estimation error. These expected advantages of the MG FPE approach are verified empirically through computer simulations in this study.

3 MG IRT FPE via the EM Algorithm for Finite Mixtures

Under the CING design for data collection, MG IRT estimation consists of estimation of both item parameters and underlying ability distributions. For such MG IRT estimation, Bock and Zimowski (1997) presented a numerical procedure suitable for an EM solution that was derived by modeling observed item response data based on a continuous ability variable and that might be approximated with a discrete version of the continuous ability variable. Their "approximate" procedure has been incorporated into the computer program BILOG-MG (Zimowski, Muraki, Mislevy, & Bock, 2003). Yet another MG IRT estimation procedure via the EM algorithm that yields essentially the same estimates as the Bock-Zimowski procedure can be derived by modeling observed

item response data based on a discrete ability variable (that is, finite mixture modeling), as done in Woodruff and Hanson (1996). The finite mixture approach for MG IRT estimation has been incorporated into the computer program ICL (Hanson, 2002), although Woodruff and Hanson (1996) presented only the EM solution for SG IRT estimation.

This section describes the essential elements of the MG IRT FPE method using the EM algorithm for finite mixtures. With notation similar to that used in Woodruff and Hanson (1996) and Kim (2006), MG IRT FPE is described as an adapted version of the usual MG IRT estimation. In both of the usual and the fixed-parameter estimation methods, the EM algorithm for finite mixtures is used to yield Bayes modal (BM) estimates for item parameters and maximum likelihood (ML) estimates for the probabilities of discrete ability distributions.

3.1 MG IRT Estimation with All Free Items

Assume the following test situation and notation. Under the common-item nonequivalent groups design, a total of J items are administered to G examinee groups and group g consists of N_g examinees. The J items may be divided into as many test forms as the number of the groups and any pair of test forms has some items in common. Examinee $i(g)$ in group g takes only the gth test form and thus the items in all other test forms are considered as "not-presented" ones. Item j has M_j response categories associated with scores $u_{j1}, \cdots, u_{jM_j}$. Denote the parameter vector for item j as $\boldsymbol{\delta}_j$, so that the parameter vector for all J items is represented by $\boldsymbol{\Delta} = (\boldsymbol{\delta}_1, \cdots, \boldsymbol{\delta}_J)$. Denote as $P_m(\theta, \boldsymbol{\delta}_j)$ the item response function for category m of item j given ability θ and the parameter vector $\boldsymbol{\delta}_j$. Then, the probability that an examinee with ability level θ earns score y_j for item j can be expressed as

$$f(y_j \mid \theta, \boldsymbol{\delta}_j) = \prod_{m=1}^{M_j} P_m(\theta, \boldsymbol{\delta}_j)^{I\{y_j = u_{jm}\}}, \tag{1}$$

where $I\{y_j = u_{jm}\}$ is equal to 1 if $y_j = u_{jm}$ and zero otherwise.

Furthermore, denote a vector of the observed item scores to all J items for examinee $i(g)$ as $\mathbf{y}_{i(g)} = (y_{i(g)1}, \cdots, y_{i(g)J})$, which includes "missing" codes such as blanks or dots. Taking the ability variable Θ as discrete, denote the latent probabilities (or weights) at K known discrete ability values $q_k (k = 1, \cdots, K)$ for group g as $\boldsymbol{\pi}_g = (\pi_{1(g)}, \cdots, \pi_{K(g)})$. Note that the same set of ability values q_k is used for all groups. The observed data likelihood for examinee i in group g can be expressed in the form of a finite mixture as follows:

$$\begin{aligned} f(\mathbf{y}_{i(g)} \mid \boldsymbol{\Delta}, \boldsymbol{\pi}_g) &= \sum_{k=1}^{K} f(\mathbf{y}_{i(g)} \mid q_k, \boldsymbol{\Delta}) \pi_{k(g)} \\ &= \sum_{k=1}^{K} \left[\prod_{j=1}^{J} f(y_{i(g)j} \mid q_k, \boldsymbol{\delta}_j) \right] \pi_{k(g)}. \end{aligned} \tag{2}$$

Consider the complete data for examinee i in group g as a joint of the observed data and the missing ability parameter $\theta_{i(g)}$, so that the complete data likelihood for the examinee may be expressed as $f(\mathbf{y}_{i(g)}, \theta_{i(g)} \mid \boldsymbol{\Delta}, \boldsymbol{\pi}_g)$. Then, based on the finite mixtures in Equation 2, the observed data likelihood for the MG sample is expressed as

$$\prod_{g=1}^{G}\prod_{i=1}^{N_g} f(\mathbf{y}_{i(g)} \mid \boldsymbol{\Delta}, \boldsymbol{\pi}_g) = \prod_{g=1}^{G}\prod_{i=1}^{N_g}\left\{\sum_{k=1}^{K} f(\mathbf{y}_{i(g)} \mid q_k, \boldsymbol{\Delta})\pi_{k(g)}\right\}. \tag{3}$$

Similarly, the complete data likelihood for the MG sample is expressed as

$$\prod_{g=1}^{G}\prod_{i=1}^{N_g} f(\mathbf{y}_{i(g)}, \theta_{i(g)} \mid \boldsymbol{\Delta}, \boldsymbol{\pi}_g) = \prod_{g=1}^{G}\prod_{i=1}^{N_g}\left\{\sum_{k=1}^{K} f(\mathbf{y}_{i(g)}, \theta_{i(g)} \mid q_k, \boldsymbol{\Delta})\pi_{k(g)}\right\}. \tag{4}$$

The BM estimates of $\boldsymbol{\Delta}$ are the values of $\boldsymbol{\Delta}$ that maximize the observed posterior distribution that is proportional to the product of the observed data likelihood and the prior for $\boldsymbol{\Delta}$, $h(\boldsymbol{\Delta})$. The EM algorithm for computing the BM estimates of $\boldsymbol{\Delta}$ uses the complete posterior distribution to compute estimates of the mode of the observed posterior distribution (Hanson, 1998; Tanner, 1996). The ML estimates of $\boldsymbol{\pi}_g (g = 1, \cdots, G)$ are the values of $\boldsymbol{\pi}_g$ that maximize the observed data likelihood, and the EM algorithm uses the complete data likelihood to find the ML estimates (Hanson, 1998; Tanner, 1996). The main task of the EM algorithm is to derive the expectation function for the parameters used in the E step and the maximization equation used in the M step. The derivation process, however, is very complicated and tedious in the case of MG IRT estimation, although the whole process is almost the same as that presented by Woodruff and Hanson (1996) except for the part dealing with $h(\boldsymbol{\Delta})$. Thus, the following describes only the final solutions of the EM algorithm without showing intermediate results of the derivation.

With the initial estimates of $\boldsymbol{\Delta}$ and $\boldsymbol{\pi}_g$, $\boldsymbol{\Delta}^{(0)}$ and $\boldsymbol{\pi}_g^{(0)}$ $(g = 1, \cdots, G)$, the expectation functions for the two parameter vectors at EM cycle $s \geq 1$ are:

$$\phi(\boldsymbol{\Delta}) = \sum_{g=1}^{G}\sum_{k=1}^{K}\sum_{j=1}^{J}\sum_{m=1}^{M_j} \log\left[P_m(q_k, \boldsymbol{\delta}_j)\right] r_{jmk(g)}^{(s-1)} + \log\left[h(\boldsymbol{\Delta})\right] \tag{5}$$

and

$$\psi(\boldsymbol{\pi}_g) = \sum_{k=1}^{K} \log\left[\pi_{k(g)}\right] n_{k(g)}^{(s-1)}, \tag{6}$$

where $r_{jmk(g)}^{(s-1)}$ is a provisional estimate of the number of examinees in the sample group g with ability value q_k who responded in category m of item j and $n_{k(g)}^{(s-1)}$ is a provisional estimate of the number of examinees with ability value q_k, and both estimates are computed by

$$r_{jmk(g)}^{(s-1)} = \sum_{i=1}^{N_g} I\left\{y_{i(g)j} = u_{jm}\right\} p(q_k \mid \mathbf{y}_{i(g)}, \boldsymbol{\Delta}^{(s-1)}, \boldsymbol{\pi}_g^{(s-1)}) \tag{7}$$

and

$$n_{k(g)}^{(s-1)} = \sum_{i=1}^{N_g} p(q_k \mid \mathbf{y}_{i(g)}, \boldsymbol{\Delta}^{(s-1)}, \boldsymbol{\pi}_g^{(s-1)}), \tag{8}$$

where $p(q_k \mid \mathbf{y}_{i(g)}, \boldsymbol{\Delta}^{(s-1)}, \boldsymbol{\pi}_g^{(s-1)})$ is the posterior probability of q_k, given $\mathbf{y}_{i(g)}$, $\boldsymbol{\Delta}^{(s-1)}$, and $\boldsymbol{\pi}_g^{(s-1)}$, and computed by

$$p(q_k \mid \mathbf{y}_{i(g)}, \boldsymbol{\Delta}^{(s-1)}, \boldsymbol{\pi}_g^{(s-1)}) = \frac{f(\mathbf{y}_{i(g)} \mid q_k, \boldsymbol{\Delta}^{(s-1)}) \pi_{k(g)}^{(s-1)}}{\sum_{k'=1}^{K} f(\mathbf{y}_{i(g)} \mid q_{k'}, \boldsymbol{\Delta}^{(s-1)}) \pi_{k'(g)}^{(s-1)}}. \tag{9}$$

In the M step at EM cycle $s \geq 1$, the BM estimates of $\boldsymbol{\Delta}$, $\boldsymbol{\Delta}^{(s)}$, are found as the values that maximize $\phi(\boldsymbol{\Delta})$ in Equation 5. Finding the BM estimates for item parameters typically involves a computer-intensive iterative technique and the details have been presented in the literature (see, e.g., Baker & Kim, 2004; Bock & Aitkin, 1981; Harwell & Baker, 1991; Mislevy, 1986; Tsutakawa & Lin, 1986). In contrast, a closed-form ML estimate of $\pi_{k(g)}$, $\pi_{k(g)}^{(s)}$, is computed by

$$\pi_{k(g)}^{(s)} = \frac{n_{k(g)}^{(s-1)}}{\sum_{k'=1}^{K} n_{k'(g)}^{(s-1)}}. \tag{10}$$

The iterative EM cycles are conducted with the provisional estimates of $\boldsymbol{\Delta}$ and $\boldsymbol{\pi}_g$ until convergence is reached. During the EM cycles, the estimates of ability distributions for examinee groups are expressed using the K known ability values q_k and the estimated weights $\boldsymbol{\pi}_g^{(s)}$. In MG IRT estimation, the scale indeterminacy is usually solved by fixing the mean and standard deviation (SD) of the ability distribution for the base group ($g = 1$) at 0 and 1, respectively. That is, the 0-1 scaling is applied to the base group, at every EM cycle or after the final EM cycle, so that all item parameters and other ability distributions are expressed on the 0-1 ability scale for the base group. However, such 0-1 scaling should not be used for FPE with MG data, as described below.

3.2 MG IRT Estimation with Some Fixed Items

To deal with FPE for MG test data, suppose that a total of J items consist of J_O "old" items whose parameters are to be fixed and J_N "new" items whose parameters are to be estimated. According to this configuration, denote the observed item scores vector for examinee $i(g)$ as $\mathbf{y}_{i(g)} = (\mathbf{y}_{Oi(g)}, \mathbf{y}_{Ni(g)})$, where $\mathbf{y}_{Oi(g)} = (y_{Oi(g)1}, \cdots, y_{Oi(g)J_O}$ and $\mathbf{y}_{Ni(g)} = (y_{Ni(g)1}, \cdots, y_{Ni(g)J_N}$. In addition, denote the parameter vector for the old items as $\boldsymbol{\Delta}_O = (\boldsymbol{\delta}_{O1}, \cdots, \boldsymbol{\delta}_{OJ_O})$ and that for the new items as $\boldsymbol{\Delta}_N = (\boldsymbol{\delta}_{N1}, \cdots, \boldsymbol{\delta}_{NJ_N})$. With the same previous notation except for the addition of subscripts N or O distinguishing new items from old items, MG IRT FPE can be conducted using the following EM algorithm.

With the initial estimates $\boldsymbol{\Delta}^{(0)}$ and $\boldsymbol{\pi}_g^{(0)}$ and the known item parameters $\boldsymbol{\Delta}_O$, the expectation functions for $\boldsymbol{\Delta}_N$ and $\boldsymbol{\pi}_g$ at EM cycle $s = 1$ are:

$$\phi(\boldsymbol{\Delta}_N) = \sum_{g=1}^{G}\sum_{k=1}^{K}\sum_{j=1}^{J_N}\sum_{m=1}^{M_j} \log\left[P_m(q_k, \boldsymbol{\delta}_{Nj})\right] r_{Njmk(g)}^{(0)} + \log\left[h(\boldsymbol{\Delta}_N)\right] \tag{11}$$

and

$$\psi(\boldsymbol{\pi}_g) = \sum_{k=1}^{K} \log\left[\pi_{k(g)}\right] n_{k(g)}^{(0)}, \tag{12}$$

where

$$r_{Njmk(g)}^{(0)} = \sum_{i=1}^{N_g} I\left\{y_{Ni(g)j} = u_{Njm}\right\} p(q_k \mid \mathbf{y}_{Oi(g)}, \boldsymbol{\Delta}_O, \boldsymbol{\pi}_g^{(0)}) \tag{13}$$

and

$$n_{k(g)}^{(0)} = \sum_{i=1}^{N_g} p(q_k \mid \mathbf{y}_{Oi(g)}, \boldsymbol{\Delta}_O, \boldsymbol{\pi}_g^{(0)}). \tag{14}$$

Note that at the first E step, with $\boldsymbol{\pi}_g^{(0)}$, only the parameters $\boldsymbol{\Delta}_O$ and observed data $\mathbf{y}_{Oi(g)}$ for the old items are used to compute the posterior probabilities of q_k . The intent of this use is that $\boldsymbol{\Delta}_N$ and $\boldsymbol{\pi}_g$ should be estimated on the scale of the old items revealed by the posterior probabilities. It should be noted that use of the known parameters $\boldsymbol{\Delta}_O$ means fixing them on the scale of the old items, which is justified by the invariance property of IRT modeling (Kim et al., 2010).

In the M step at EM cycle $s = 1$, the first BM estimates of $\boldsymbol{\Delta}_N$, $\boldsymbol{\Delta}_N^{(1)}$, are found as the values that maximize $\phi(\boldsymbol{\Delta}_N)$ in Equation 11. In addition, based on Equation 14, the first ML estimate of $\pi_{k(g)}$, $\pi_{k(g)}^{(1)}$, is computed for all examinee groups by

$$\pi_{k(g)}^{(1)} = \frac{n_{k(g)}^{(0)}}{\sum_{k'=1}^{K} n_{k'(g)}^{(0)}}. \tag{15}$$

However, at EM cycles $s \geq 2$, the expectation functions for $\boldsymbol{\Delta}_N$ and $\boldsymbol{\pi}_g$ are modified into:

$$\phi(\boldsymbol{\Delta}_N) = \sum_{g=1}^{G}\sum_{k=1}^{K}\sum_{j=1}^{J_N}\sum_{m=1}^{M_j} \log\left[P_m(q_k, \boldsymbol{\delta}_{Nj})\right] r_{Njmk(g)}^{(s-1)} + \log\left[h(\boldsymbol{\Delta}_N)\right] \tag{16}$$

and

$$\psi(\boldsymbol{\pi}_g) = \sum_{k=1}^{K} \log\left[\pi_{k(g)}\right] n_{k(g)}^{(s-1)}, \tag{17}$$

where

$$r_{Njmk(g)}^{(s-1)} = \sum_{i=1}^{N_g} I\left\{y_{Ni(g)j} = u_{Njm}\right\} p(q_k \mid \mathbf{y}_{Oi(g)}, \boldsymbol{\Delta}_O, \boldsymbol{\Delta}_N^{(s-1)}, \boldsymbol{\pi}_g^{(s-1)}) \tag{18}$$

and

$$n_{k(g)}^{(s-1)} = \sum_{i=1}^{N_g} p(q_k \mid \mathbf{y}_{Oi(g)}, \boldsymbol{\Delta}_O, \boldsymbol{\Delta}_N^{(s-1)}, \boldsymbol{\pi}_g^{(s-1)}). \tag{19}$$

Note the modification for EM cycles $s \geq 2$ that the posterior probabilities of q_k are computed based on all the observed data for both old and new items $\mathbf{y}_{i(g)}$, the known fixed item parameters $\boldsymbol{\Delta}_O$, the estimated item parameters $\boldsymbol{\Delta}_N^{(s-1)}$, and the estimated weights $\boldsymbol{\pi}_g^{(s-1)}$. The logic for this modification is that the underlying ability distributions for G groups would be more precisely estimated (i.e., recovered) when more items (old + new) are used for the estimation, as long as the chosen IRT model fits the response data for the items (Kim, 2006).

In the M step at EM cycle $s \geq 2$, the BM estimates of $\boldsymbol{\Delta}_N$, $\boldsymbol{\Delta}_N^{(s)}$, are found as the values that maximize $\phi(\boldsymbol{\Delta}_N)$ in Equation 16. Based on Equation 19, the ML estimate of $\pi_{k(g)}$, $\pi_{k(g)}^{(s)}$, is computed by

$$\pi_{k(g)}^{(s)} = \frac{n_{k(g)}^{(s-1)}}{\sum_{k'=1}^{K} n_{k'(g)}^{(s-1)}}, \tag{20}$$

which is in form the same as Equation 10.

As for the usual MG IRT estimation, the iterative EM cycles for FPE are conducted with the provisional estimates of $\boldsymbol{\Delta}_N$ and $\boldsymbol{\pi}_g$ until convergence is reached. However, during the EM cycles, no 0-1 scaling is used for the base group to solve IRT scale indeterminacy, because the already established scale of the old items is used to estimate the parameters of the new items and the underlying ability distributions for all examinee groups. All that is needed is to perform the EM cycles iteratively, with reasonable initial estimates of $\boldsymbol{\Delta}_N$ and $\boldsymbol{\pi}_g$. Note that a successful FPE run can be accomplished through the estimation of the underlying ability distributions for all the groups, including the base group.

4 Simulation Methods

A simulation study was conducted to investigate the performance of the FPE method extended for MG test data. The performance of the MG FPE method was compared with those of the "MG 0-1 estimation and linking" method and the "SG 0-1 estimation and linking" method described earlier in this paper, in recovering the true parameters of items and the underlying ability distributions. The details of simulation methods are as follows.

4.1 Design and the Parameters for Simulation

Test data were simulated under the CING design to examine IRT estimation with MG while maintaining the established ability scale (θ_0). As illustrated in Figure 1, the CING design involved two examinee groups (Groups 1 and 2) and

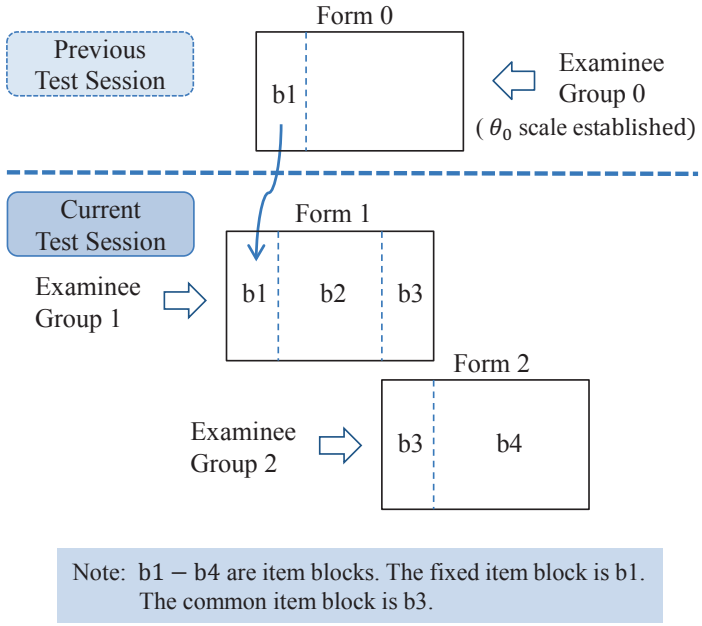

Figure 1: The CING design involving 2 groups and 2 forms used for simulation study

two test forms (Forms 1 and 2), and Form 1 was connected with the base form, Form 0, whose data were to be analyzed using the θ_0 scale.

The following was assumed for the CING design. All test forms have the same number (40) of dichotomously-scored items. All items are to be analyzed by the three-parameter logistic (3PL) model (Lord, 1980), whose response function for the "correct" category of item j is expressed as $P[\theta, \boldsymbol{\delta}_j = (a_j, b_j, c_j)] = c_j + (1 - c_j)/\{1 + \exp[-1.7a_j(\theta - b_j)]\}$, where a_j, b_j, and c_j are the discrimination, difficulty, and pseudo-guessing parameters, respectively. Forms 1 and 2 have 10 items (the b3 block in Figure 1) in common. Form 1 consists of 10 "old" items and 30 "new" items. The 10 old items of Form 1 come from Form 0, administered previously, and their parameters (estimated) are already established on the θ_0 scale. The 10 old items play the role of an anchor that links the current MG test data with the previous test data. All items of Form 2 are "new" items. Forms 1 and 2 are administered to the two nonequivalent groups, Groups 1 and 2, respectively. Once MG test data are obtained, the task of interest is to estimate the new item parameters and the underlying ability distributions for the groups on the scale, not on an arbitrary 0-1 scale.

To produce item parameters with realistic properties for the three forms (Forms 0 through 2), 100 items with 3PL parameters were first chosen from the science assessment presented in the 1996 National Assessment of Educational

Progress report (Allen, Carlson, & Zelenak, 1999). The 100 items were divided into ten sets of 10 items, such that all item sets might be as similar as possible to one another in terms of the distributions of the a, b, and c parameters. Four of the ten sets were assigned to Form 0. For simplicity, the first 10 items of Form 0 were designated as the anchor linking Form 0 with Form 1. The 10 linking items and 30 "new" items (the b2 and b3 blocks) from other three 10-item sets comprised Form 1. The last 10 items of Form 1 and 30 "new" items (the b4 block) comprised Form 2, so that Forms 1 and 2 might have the 10 items (the b3 block) in common. Consequently, the three forms were very similar to each other in terms of distributions of the 3PL item parameters. For the combined parameter set of all the forms, a parameters ranged from 0.38 to 1.61 with means of 0.74 to 0.77; b parameters ranged from -2.54 to 2.74 with means of 0.21 to 0.22; and c parameters ranged from 0.10 to 0.34 with means of 0.22 to 0.23.

The underlying ability distributions for examinee groups were assumed as follows. Denoting $\mathrm{N}(\mu,\ \sigma^2)$ as a normal distribution having a mean of μ and an SD of σ, Group 0 taking Form 0 has a N(0, 1) ability distribution. The 0-1 scale defined by Group 0 is used as the established scale. Groups 1 and 2 have $\mathrm{N}(-0.1,\ 0.8^2)$ and $\mathrm{N}(0.5,\ 1.2^2)$ ability distributions, respectively, so the mean ability levels of the two groups are slightly lower or higher than the average difficulties of the three forms.

The following sample sizes were assumed for each of the examinee groups. Form 0, administered to Group 0, was used to obtain "realistic" parameter estimates for the 10 linking items to be fixed, because the linking items' parameters are usually unknown in practice. Kim (2006) found that the item-parameter recovery with FPE is robust to sampling error in estimates of the fixed items, whether they were analyzed with 300 examinees or with 3,000 examinees. Taking this point into account, the sample size for Group 0 was set at 1,000, a suitable size leading to stable 3PL item parameter estimates. On the other hand, two sample sizes, 500 and 2,000, were chosen for the two groups, Groups 1 and 2, to examine the performance of the three IRT estimation methods using a small and a large sample size.

4.2 Data Generation and IRT Estimation and Linking

In each of the two sample-size (500 and 2,000) conditions, 100 replications of SG or MG data generation and IRT estimation and linking were carried out to investigate the performance of the MG FPE method and the other two methods. In each replication, the data matrices for Forms 0 through 2 were generated using the item parameters ($\boldsymbol{\delta}_j$) on the forms and the ability (θ) parameters randomly sampled from the normal distributions for Groups 0 through 2. Item scores (0/1) in the matrices were generated using the rule, as used in Kim (2006): if $P(\theta, \boldsymbol{\delta}_j) \geq R$ then 1, otherwise 0, where R is a random number from a uniform distribution, ranging from 0 to 1.

Different input formats were used for the generation of SG and MG data matrices. To implement SG 0-1 estimation, a data matrix of "sample-size $\times$ 40"

was generated for each of Forms 0 through 2, including no missing codes for item scores. To conduct MG 0-1 estimation or MG FPE, an augmented data matrix of "sample-size$\times$70", including the missing codes (blanks) for the 30 not-presented items, was generated for each form and the two data matrices were combined into one. Especially, the codes of "group indicator" were put in all of MG test data, so that MG IRT estimation could be implemented appropriately.

Prior to implementing the three methods in each replication, the processes of separate 0-1 estimation and linking were conducted for the SG data of Form 0 to obtain the parameter estimates of the 10 linking items that were placed on the θ_0 scale. The separate 0-1 estimation was carried out using ICL, with 51 points of ability (equally spaced from -5 to 5) following a N(0, 1) distribution, with 200 EM cycles, and with the convergence criterion of 0.001. In addition, for BM estimation of all items, the priors for a and c parameters were specified, respectively, as a lognormal distribution having mean = 0 and SD = 0.5 on the logarithm and a beta distribution with two parameters $\alpha = 5$ and $\beta = 17$, where the beta density function $f(c)$ is proportional to $c^{\alpha-1}(1-c)^{\beta-1}$. Note that the two prior distributions are the default ones that are supplied by BILOG-MG (R. J. Mislevy, personal communication, July 11, 2013; he confirmed that BILOG-MG's ALPHA = α and BETA = β). Posterior weights (i.e., probabilities) at the 51 ability points were computed based on the underlying ability distribution and the converged item parameter estimates. The resulting posterior ability distribution and item parameter estimates were not intended to be on the θ_0 scale (Kim, 2006). The linking coefficients, A and B, from applying the Stocking-Lord (Stocking & Lord, 1983) method to the 10 linking items (having two sets of true parameters and their estimates) were used to transform the ability points and item parameter estimates onto the θ_0 scale. Specifically, with the transformed values starred (*), the ability points were transformed by the relation $\theta_q^* = A\theta_q + B$; the 3PL parameter estimates were transformed by the relations, $\hat{a}_j^* = \hat{a}_j/A$, $\hat{b}_j^* = A\hat{b}_j + B$, and $\hat{c}_j^* = \hat{c}_j$. To implement the Stocking-Lord method, the computer program STUIRT (Kim & Kolen, 2004) was used.

Along with the parameter estimates of the linking items for Form 0 and the posterior ability distribution for Group 0, placed on the θ_0 scale, the three IRT estimation methods to maintain the established ability scale (θ_0) were implemented in turn, as follows. First, the MG FPE method was implemented using a computer program that was created for this study, based on the C++ code developed by Hanson (2002) and used for ICL. The implementation can also be achieved by using ICL's functions, and a command file for such implementation is available upon request from the authors. For MG FPE, the parameters of the first 10 linking items on Form 1 were fixed at the values of the estimates on the θ_0 scale and 51 points of ability (equally spaced from -5 to 5) were used for all groups. The settings for EM cycles, convergence criterion, and Bayesian priors were the same as those used for the SG 0-1 estimation with Form 0. Second, the "MG 0-1 estimation and linking" method was implemented using the ICL program and the Stocking-Lord linking method. The MG 0-1 estimation was

conducted with Group 1 being designated as the base group, which defined an arbitrary 0-1 scale (θ_1). The estimates of new items' parameters and ability distributions from MG 0-1 estimation were transformed to the θ_0 scale using the Stocking-Lord method that produced the "θ_1 to θ_0" linking coefficients based on the two sets of parameter estimates for the 10 linking items.

Third, the "SG 0-1 estimation and linking" method was conducted using the ICL program and the Stocking-Lord method. For this method, separate 0-1 estimation was conducted for each form and then one or two scale linkings, based on the linking or common items between forms, were done to link each of the θ_1 and θ_2 scales from Groups 1 and 2 to the θ_0 scale. Denote the linking coefficients of a "θ_g to θ_{g-1}" linear transformation as A_g and B_g. For Form 1, the posterior ability distribution and item parameter estimates were transformed onto the θ_0 scale by the relation $\theta_0 = A_1\theta_1 + B_1$. For Form 2, the chained relation $\theta_0 = (A_1A_2)\theta_2 + (B_1 + A_1B_2)$ was used for scale linking.

4.3 Evaluation of the Methods

The performance of the three IRT estimation methods over 100 replications was evaluated by assessing how well the underlying ability distributions and the parameters for non-fixed new items were recovered. Basically, the mean and mean squared error (MSE) of parameter estimates from replications were computed and used as evaluation criteria.

To evaluate the recovery of the underlying ability distribution for a group, the mean and SD of the posterior ability distribution were computed for each method in each replication. With the resulting 100 estimated sets of means and SDs for a group, the mean and root of mean squared error (RMSE) were computed for each of the means and SDs. Note that the underlying ability distributions for Groups 1 and 2 were N(-0.1, 0.8^2) and N(0.5, 1.2^2), respectively.

For each of Forms 1 and 2, item-parameter recovery was evaluated only for the non-fixed new items. The non-fixed new items were divided into two item blocks, common and non-common. Compared to the non-common item block, the common item block is associated with a doubled sample size and thus is expected to have less estimation error on the average. Considering this point, the recovery of each of the 3PL item parameters was evaluated by item block. Specifically, for each of the a, b, and c parameters, the MSE of parameter estimates was computed for every each item in a block and the average of the MSE values over the block items was computed. The root of the average MSE was computed and simply referred to as RMSE (specifically, a-RMSE, b-RMSE, and c-RMSE) for each parameter.

5 Results

The recovery results for the true mean and SD of each underlying ability distribution were first examined. As seen from Table 1, for both sample sizes, the three methods produced the means of estimated means and SDs for each

Table 1: Mean and RMSE of Means and SDs of Estimated Ability Distributions by Group and Sample Size (SS)

			Mean		SD	
			Mean	RMSE	Mean	RMSE
Group 1			(−0.1)		(0.8)	
	SS = 500					
		MG FPE	−0.10	0.049	0.82	0.047
		MG 0-1 & L.	−0.10	0.052	0.83	0.063
		SG 0-1 & L.	−0.11	0.056	0.85	0.069
	SS = 2,000					
		MG FPE	−0.12	0.034	0.81	0.040
		MG 0-1 & L.	−0.12	0.036	0.81	0.044
		SG 0-1 & L.	−0.12	0.040	0.82	0.043
Group 2			(0.5)		(1.2)	
	SS = 500					
		MG FPE	0.50	0.074	1.22	0.076
		MG 0-1 & L.	0.51	0.078	1.24	0.127
		SG 0-1 & L.	0.51	0.086	1.23	0.135
	SS = 2,000					
		MG FPE	0.48	0.048	1.22	0.058
		MG 0-1 & L.	0.48	0.053	1.20	0.072
		SG 0-1 & L.	0.48	0.050	1.20	0.079

Note. The values in parentheses are the true means and standard deviations of underlying ability distributions for each examinee group.

group that were close to the true mean and SD of the underlying ability distribution. In addition, as expected, the values of RMSE of estimated means and SDs decreased when the sample size per form increased. These results suggest that the three methods performed well for the SG or MG test data and that the underlying ability distributions were properly recovered on the established scale.

However, the three methods showed somewhat consistent, non-negligible differences on the RMSE statistic. For mean-RMSE and SD-RMSE, the MG FPE method produced the smallest values in all cases, and with a few exceptions the "MG 0-1 estimation and linking" method produced the second smallest values and the "SG 0-1 estimation and linking" method produced the largest values. These results indicate that the MG FPE method was more stable and accurate for recovering the underlying ability distributions than the other two methods.

The item-parameter recovery was examined by form, item block, and sam-

Table 2: RMSE of 3PL Item Parameter Estimates for Form 1 by Item Block and Sample Size (SS)

	20 Non-Common Items (No. 11–30)			10 Common Items (No. 31–40)		
	a-RMSE	*b*-RMSE	*c*-RMSE	*a*-RMSE	*b*-RMSE	*c*-RMSE
SS = 500						
MG FPE	0.209	0.320	0.040	0.131	0.222	0.041
MG 0-1 & L.	0.214	0.356	0.041	0.134	0.248	0.041
SG 0-1 & L.	0.230	0.324	0.039	0.191	0.306	0.040
SS = 2,000						
MG FPE	0.150	0.218	0.038	0.079	0.152	0.038
MG 0-1 & L.	0.152	0.232	0.039	0.085	0.160	0.038
SG 0-1 & L.	0.168	0.225	0.039	0.147	0.210	0.037

ple size. First, the results for Form 1 are presented in Table 2. For Form 1, as expected for all the methods, the values of all RMSE decreased as the sample size increased. For the four conditions of two sample sizes by two item blocks (common items vs. non-common items), the three methods produced very similar values of *c*-RMSE. However, for *a*-RMSE and *b*-RMSE, the relative performance of the three methods depended on the item block. For the block of 20 non-common items (No. 11–30), the MG FPE method produced the smallest values of *a*-RMSE and *b*-RMSE for both sample sizes. For both sample sizes, the "MG 0-1 estimation and linking" method produced the second smallest values of *a*-RMSE but produced the largest values of *b*-RMSE. For the common item block, for both sample sizes, the MG FPE method produced the smallest values of *a*-RMSE and *b*-RMSE, and the "MG 0-1 estimation and linking" method produced the second smallest values for those criteria. A noteworthy point was that the differences in RMSE between the MG FPE method and the "SG 0-1 estimation and linking" method were larger for the common item block than for the non-common item block.

The results for Form 2 are presented in Table 3. As for Form 1, for all the methods, the values of all RMSE statistics decreased as the sample size increased, which is as expected. Again, as for Form 1, the three methods produced very similar values of *c*-RMSE, regardless of sample size and item block type. However, for *a*-RMSE and *b*-RMSE, the three methods showed substantial differences across conditions. For the common item block with both sample sizes, the MG FPE method produced the smallest values of *a*-RMSE and *b*-RMSE and the "MG 0-1 estimation and linking" method produced the second smallest values for those criteria. It should be noted that each of the two "MG estima-

Table 3: RMSE of 3PL Item Parameter Estimates for Form 2 by Item Block and Sample Size (SS)

	10 Common Items (No. 1–10)			30 Non-Common Items (No. 11–40)		
	a-RMSE	*b*-RMSE	*c*-RMSE	*a*-RMSE	*b*-RMSE	*c*-RMSE
SS = 500						
MG FPE	0.131	0.222	0.041	0.145	0.213	0.045
MG 0-1 & L.	0.134	0.248	0.041	0.149	0.228	0.045
SG 0-1 & L.	0.147	0.295	0.043	0.157	0.237	0.044
SS = 2,000						
MG FPE	0.079	0.152	0.038	0.087	0.161	0.042
MG 0-1 & L.	0.085	0.160	0.038	0.092	0.163	0.042
SG 0-1 & L.	0.099	0.176	0.038	0.095	0.165	0.041

tion" methods produced one set of results for the common items and thus the values of all RMSE for the two methods presented in Table 3 are the same as those presented in Table 2. Similarly, for the non-common item block with both sample sizes, the MG FPE method produced the smallest values of a-RMSE and b-RMSE and the "SG 0-1 estimation and linking" method produced the largest values for those criteria. As for Form 1, it was noteworthy that the differences in RMSE between the MG FPE method and the "SG 0-1 estimation and linking" method were larger for the common item block than for the non-common item block.

6 Discussion

The results of the simulation conducted under the CING design show that, overall, the MG FPE method performed best (had the least error) for estimating item parameters and ability distributions on the established scale. With a few exceptions, the "MG 0-1 estimation and linking" method was the second best performer and the "SG 0-1 estimation and linking" method was the poorest performer. Compared to the "SG 0-1 estimation and linking" method, the better performance of the MG FPE method was more noteworthy for the common items between forms than for the non-common items.

The relative performance of the three methods in estimating the item parameters on the established scale can be explained as follows. Parameter estimation for non-common items depends largely on two related factors: (a) the degree of recovery of underlying ability distributions on the established scale and (b) the magnitude of scale linking error. The MG FPE method performed best in item-

parameter estimation, likely because all ability distributions were well estimated on the established scale and no scale linking was used. The "MG 0-1 estimation and linking" method tended to perform second best in item-parameter estimation, likely because all ability distributions were concurrently estimated on a 0-1 scale, but transformed once to the established scale, so that linking error contributed to error in the item parameter estimates. The "SG 0-1 estimation and linking" method tended to perform the poorest, likely because ability distributions were separately estimated on each 0-1 scale and transformed, one or more times, to the established scale so that accumulated linking error contributed to error in the item parameter estimates.

On the other hand, parameter estimation for common items likely depends on three factors, including sample size. For the common items, the two "MG estimation" methods use a larger sample size than the "SG estimation" method for item-parameter estimation. The two "MG estimation" methods tended to perform better than the "SG estimation" method, mainly because a larger sample size leads to less error in item-parameter estimation. In addition, the relative performance of the three methods also can be explained by two other factors: (a) the degree of recovery of the underlying ability distributions and (b) the magnitude of scale linking error.

The preceding discussion suggests that the appropriate recovery of the underlying ability distributions is critical for MG FPE, as is the case for SG FPE (Kim, 2006). Related to this point, a major concern is that use of inaccurate parameter estimates of the old items that are fixed could possibly lead to poor estimation of the underlying ability distributions. This concern is alleviated partly by the technical mechanism of the MG FPE method. Likely the old items to be fixed play a central role in recovering the underlying ability distributions on the established scale. However, the data for the other new items are also used in the process of FPE so the underlying ability distributions are estimated onto the established scale through both data of the new and old items (see Equations 18, 19, and 20). Thus, adverse effects of "bad" old item parameter estimates on FPE might not be a substantial concern, unless many of the parameters for the old items are poorly estimated on the established scale.

Practical measures can also be taken to control the possible effects of some poor old item parameter estimates on FPE. Before conducting FPE, SG 0-1 estimation can be implemented using the data from the new form that includes both the old and new items. Then, two sets of item parameter estimates for the old items to be fixed are plotted to look for "outlier" old items whose item parameter estimates do not appear to lie on a straight line. The old items that are judged to be outliers are not used as the linking items for FPE. This process of finding and eliminating bad linking items is typically considered in IRT scale linking (see Kolen & Brennan, 2014, chap. 6, for an example).

Further, at least three conditions should be considered to make the best use of the MG FPE. First, the method is conditional on the fit of model(s) for analysis to test data. In fact, this "fit" condition applies to any IRT estimation method. When an acceptable degree of model-to-data fit is not achieved, the meaning of the IRT estimates is questionable. Second, the IRT invariance prop-

erty holds for all test items. The invariance (within a linear transformation) of item parameters will hold with simulated test data that fit the IRT model, but cannot be guaranteed in practice. Note that the issue of "bad" old item parameters to be fixed is partly related to this invariance property. Invariance, especially for the fixed items, may be threatened by several factors such as item exposure, examinee motivation, and context effects. Third, the common-item link between test forms is strong enough for the MG estimation to be properly conducted. The common items should be administered to all the examinees in each group, not to a subset of examinees. With these conditions being met, the MG FPE method is expected to perform better when there are more fixed items, more common items between forms, and larger sample sizes. However, little is known regarding how the performance of the FPE method is affected by violations of such conditions. Further research is warranted that addresses such practical issues.

7 References

Allen, N. L., Carlson, J. E., & Zelenak, C. A. (1999). *The NAEP 1996 technical report* (NCES 1999-452). Washington, DC: National Center for Education Statistics.

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

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

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

Bock, R. D., & Zimowski, M. F. (1997). Multiple group IRT. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 433–448). New York: Springer.

de Ayala, R. J. (2009). *The theory and practice of item response theory*. New York: Guilford.

DeMars, C. E., & Jurich, D. P. (2012). Software note: Using BILOG for fixed-anchor item calibration. *Applied Psychological Measurement, 36*, 232–236.

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

Divgi, D. R. (1985). A minimum chi-square method for developing a common metric in item response theory. *Applied Psychological Measurement, 9*, 413–415.

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

Hambleton, R. K. (1989). Principles and selected applications of item response theory. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 147–200). New York: American Council on Education and Macmillan.

Hanson, B. A. (1998, April). *Bayes modal estimates of a discrete latent variable distribution in item response theory models using the EM algorithm.* Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.

Hanson, B. A. (2002). IRT Command Language (Version 0.020301). Monterey, CA: Author.

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.

Harwell, M. R., & Baker, F. B. (1991). The use of prior distributions in marginalized Bayesian item parameter estimation: A didactic. *Applied Psychological Measurement, 15*, 375–389.

Keller, L. A., & Hambleton, R. K. (2013). The long-term sustainability of IRT scaling methods in mixed-format tests. *Journal of Educational Measurement, 50*, 390–407.

Keller, L. A., & Keller, R. R. (2011). The long-term sustainability of different item response theory scaling methods. *Educational and Psychological Measurement, 71*, 362–379.

Kim, S. (2006). A comparative study of IRT fixed parameter calibration methods. *Journal of Educational Measurement, 43*, 353–381.

Kim, S., Harris, D. J., & Kolen, M. J. (2010). Equating with polytomous item response models. In M. L. Nering & R. Ostini (Eds.), *Handbook of polytomous item response theory models* (pp. 257–291). New York: Routledge.

Kim, S., & Kolen, M. J. (2004). STUIRT [Computer software]. Iowa City, IA: Iowa Testing Programs, University of Iowa. Available from http://www.education.uiowa.edu/casma

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.

Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scaling, and linking: Methods and practices* (3rd ed.). New York: Springer.

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

Loyd, B. H., & Hoover, H. D. (1980). Vertical equating using the Rasch model. *Journal of Educational Measurement, 17*, 179–193.

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

McLachlan, G. J., & Peel, D. (2000). *Finite mixture models*. New York: John Wiley & Sons.

Mislevy, R. J. (1986). Bayes modal estimation in item response models. *Psychometrika, 51*, 177–195.

Mislevy, R. J., & Bock, R. D. (1990). *BILOG 3: Item analysis and test scoring with binary logistic models* (2nd ed.). Mooresville, IN: Scientific Software International.

Ogasawara, H. (2001). Least squares estimation of item response theory linking coefficients. *Applied Psychological Measurement, 25*, 373–383.

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

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

Tanner, M. (1996). *Tools for statistical inference: Methods for the exploration of posterior distributions and likelihood functions* (3rd ed.). New York: Springer-Verlag.

Titterington, D. M., Smith, A. F. M., & Makov, U. E. (1985). *Statistical analysis of finite mixture distributions*. New York: John Wiley & Sons.

Tsutakawa, R. K., & Lin, H. Y. (1986). Bayesian estimation of item response curves. *Psychometrika, 51*, 251–267.

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

Woodruff, D. J., & Hanson, B. A. (1996). *Estimation of item response models using the EM algorithm for finite mixtures* (ACT Research Report 96-6). Iowa City, IA: ACT, Inc.

Yen, W. M., & Fitzpatrick, A. R. (2006). Item response theory. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 111–153). Westport, CT: American Council on Education and Praeger.

Young, M. J. (2006). Vertical scales. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 469–485). Mahwah, NJ: Lawrence Erlbaum Associates.

Zimowski, M. F., Muraki, E., Mislevy, R. J., & Bock, R. D. (2003). BILOG-MG: Multiple-group IRT analysis and test maintenance for binary items [Computer software]. Chicago, IL: Scientific Software International.