LINKING TESTS UNDER THE CONTINUOUS RESPONSE MODEL

Kojiro Shojima*,**

In this study, after defining the equating coefficients of the continuous response model (CRM, Samejima, 1973, 1974), we proposed three procedures of linking tests under the CRM in the context of both common examinees and items designs. One was for the common examinees design, and the other two were for the common items design. As for the common examinees design, we proposed a method for estimating the equating coefficients using the marginal maximum likelihood estimation with the EM algorithm, where each common examinee's latent trait $\theta$, which becomes a nuisance parameter, was integrated over the posterior distribution of $\theta$. Under the common items design, we applied the weighted least squares method (Haebara, 1980) and the test characteristic curve method (Stocking & Lord, 1983) to the CRM after introducing the item response function of the CRM. We also confirmed the accuracy of the three proposed methods using simulation data and actual data.

1. INTRODUCTION

Item response theory (IRT, Lord, 1980) has become the most powerful tool in examining both subjects' latent traits and the characteristics of a test. In order to utilize tests over a long period of time, equating/linking is one of the most important tasks in IRT, test theory and other psychometric measurement theories for comparing several tests on the same scale.

A principal concern in linking is to estimate the equating coefficients, or equating constants (Hambleton & Swaminathan, 1985, p.205). There are roughly two positions in linking/equating: common examinees and items design. Common examinees design is a linking plan which uses the fact that the $\theta$ metric is unique for all common examinees through different tests. In contrast, the common items design makes use of common items' information, common items being items adopted in two or more tests. Common items design is based on the logic that different item parameter values estimated from different data sets of the same item are fundamentally equivalent.

In recent years, many attempts to apply IRT to psychological questionnaires or mental tests have been made. Most of these tests consist of Likert-type items. IRT models applicable to Likert-type data are the graded response model (GRM, Samejima, 1969), partial credit model (PCM, Masters, 1982) and generalized PCM (GPCM, Muraki, 1992). However, $K - 1$ difficulty parameters are required for each item when the item has $K$ categories under GRM, PCM and GPCM. Therefore, when $K$ is large, stable estimates cannot be obtained without a sufficient number of observations.

Key Words and Phrases: item response theory, linking, equating coefficients, common examinees design, common items design, continuous response model, marginal maximum likelihood estimation, EM algorithm, weighted least squares method, test characteristic curve method.

* Research Division, The National Center for University Entrance Examinations, 2-19-23 Komaba Meguro-ku, Tokyo 153-8501 JAPAN
** Graduate School of Letters, Arts and Sciences, Waseda University. Home address: 6-4-205 Irifune Urayasu, Chiba, Japan 279-0012. E-mail: shojima@rd.dnc.ac.jp
Most items in psychological questionnaires have five or seven categories. In order to apply GRM, PCM and GPCM to such items, four difficulty parameters are needed with five-category items, and six difficulty parameters are required for seven-category items. Therefore, to acquire small errors in estimation, thousands or tens of thousands of observations must be secured, which is difficult in small-scale research.

Samejima (1973, 1974) proposed the continuous response model (CRM). The CRM is generalized as a limiting form of GRM and can treat item scores as continuous values. The CRM assumes three parameters per item without the guessing parameter, one of them being the discrimination parameter, and the number of parameters specifying the item difficulty is always two, regardless of $K$.

Since Likert-type data is, strictly speaking, rank-ordered, it is logically correct to apply GRM and GPCM to this type of data. However, researches regarding Likert-type data as continuous scores and applying factor analysis are very popular, and there seems to be little reluctance to stand down on this viewpoint. Above all, the number of parameters of the CRM is always fixed with any $K$, and this is considered to be desirable, especially for psychological questionnaires with a comparatively large $K$.

Moreover, although grading was difficult with paper-pencil tests, computer-based testing enables subjects to check anywhere on the axis of “extremely positive” to “extremely negative” (see Samejima, 1973, p.204, Figure 1). It can be said that the CRM is most suitable for such items.

Wang and Zeng (1998) discussed the estimation method of the CRM item parameters using the marginal maximum likelihood estimation with an EM algorithm (Dempster, Laird & Rubin, 1977). The next step is to equate/link tests under the CRM, or to demonstrate a method to enrich the item pool under the CRM. The main purpose of this study is to discuss the linking method when items are calibrated by the CRM. Specifically, by extending the ideas of Noguchi (1986) and Ogawara (2001a), we propose a method for estimating equating coefficients using the marginal maximum likelihood estimation with the EM algorithm (MML–EM) under the common examinees design. As for the common items design, we applied the weighted least squares method (Haebara, 1980) and the test characteristic curve method (Stocking & Lord, 1983) directly to the CRM after introducing the item response function of the CRM. The second purpose of our study is to confirm the accuracy of the proposed methods using simulation data and actual data.

2. METHOD

2.1 The Continuous Response Model

The probability of an examinee with a trait level $\theta$ obtaining a score $x_j$ on a score scale ranging from 0 to $k_j$ for an item $j$ is assumed to be

$$\Pr(X_j = x_j | \theta) := P_{jx_j}(\theta) = \frac{a_j}{\sqrt{2\pi}\alpha_j} \exp\left\{-\frac{1}{2} a_j^2 \left( \frac{\theta - b_j - \frac{1}{\alpha_j} \ln \frac{x_j}{k_j-x_j}}{\alpha_j} \right)^2 \right\}$$

$$= \frac{a_j}{\sqrt{2\pi}\alpha_j} \exp\left\{-\frac{1}{2} a_j^2 \left( \theta - b_j - \frac{x_j}{\alpha_j} \right)^2 \right\}$$
in normal density type CRM (Samejima, 1973), where \( z_j = \ln \frac{x_j}{(k_j - x_j)} \) \((-\infty < z_j < \infty\)). To simplify the algebraic operations, \( z_j \) is henceforth used instead of \( x_j \). The probability of an examinee with a specific \( \theta \) obtaining a score \( z_j \) or higher on an item \( j \) is defined as
\[
\Pr(Z_j \geq z_j | \theta) := P_{z_j}^d (\theta) = \Phi_{z_j} (\theta - b_j - z_j / \alpha_j) \quad \text{if } z_j < -\infty
\]
where \( N(-|0,1) \) is the probability density function of the standard normal.

The parameters \( a_j (>0) \) and \( b_j (-\infty, \infty) \) are the discrimination and the difficulty of item \( j \), respectively. The parameter \( \alpha_j (>0) \) is a scaling parameter that defines some scale transformation linking the original rating score scale to the \( \theta \) scale (Wang & Zeng, 1998). Henceforth, \( P_{z_j}^d (\theta) \) will be written as \( P_{z_j}^d (\theta; a_j, b_j, \alpha_j) \).

2.2 Scale Indeterminacy and Equating Coefficients of the CRM

Using \( h \) and \( l \) whose domains are \( 0 < h < \infty \) and \( -\infty < l < \infty \), respectively, and assuming the following linear transformations,
\[
\theta^* = h\theta + l \\
a_j^* = a_j / h \\
b_j^* = h b_j + l \\
\alpha_j^* = \alpha_j / h,
\]
we obtain
\[
P_{z_j}^d (\theta^*; a_j^*, b_j^*, \alpha_j^*) = \frac{a_j^*}{\sqrt{2\pi}} \exp \left\{ -\frac{1}{2} \frac{a_j^*}{\beta_j^*} (\theta^* - b_j^* - z_j^*)^2 \right\}.
\]
The CRM also has a problem of arbitrariness of origin and scale on the latent \( \theta \) axis as do many other IRT models. Two scales can be linked if \( h \) and \( l \) are known, as described by Lord (1980). However, since \( h \) and \( l \) are unknown in almost all cases, the work of equating becomes an estimation problem of \( h \) and \( l \) as a result. The constants \( h \) and \( l \) are called equating coefficients.

2.3 Estimation of Equating Coefficients

2.3.1 Common Examinees Design

Equating methods in the context of the common examinees design have been discussed by Marco (1977), Noguchi (1983, 1986, 1990), Toyoda (1986) and Ogasawara (2001a) in the case of dichotomous response models. In this study, we developed the methods proposed by Noguchi (1986) and Ogasawara (2001a) and applied them to the CRM. Suppose that there are \( N \) common examinees, and let \( \theta = [\theta_1, \cdots, \theta_N]' \) represent the vector of latent \( \theta \) random variables for \( N \) examinees. Let \( Z_B = \{z_{Bij}\} \) represent an \( N \times n_B \)
matrix of the response data of $N$ common examinees for $n_B$ items on the base test, and $z_{Bij} = \ln x_{Bij}/(k_j - x_{Bij})$ represents the response of person $i$ for an item $j$. Similarly, let $Z_T$ be an $N \times n_T$ matrix of the response data of $N$ common examinees on $n_T$ items of the target test. The variables $n_B$ and $n_T$ are the test length of the base and the target test, respectively. Let us also suppose that the estimated results of item parameters from the base test and the target test are $\hat{\bar{\xi}}_B = \{\hat{\xi}_{Bij}\}$ and $\hat{\bar{\xi}}_T = \{\hat{\xi}_{Tik}\}$, respectively, where $\hat{\xi}_{Bij} = (\hat{a}_{Bj} \hat{b}_{Bj} \hat{\alpha}_{Bj})'(j = 1, 2, \ldots, n_B)$ and $\hat{\xi}_{Tik} = (\hat{a}_{Tk} \hat{b}_{Tk} \hat{\alpha}_{Tk})'(k = 1, 2, \ldots, n_T)$.

Assuming that $\hat{\bar{\xi}}_T = \{\hat{\xi}_{Tik}\}$ is item parameter set $\hat{\bar{\xi}}_T$ equated onto the base test scale, the likelihood function of the response matrices $Z_B, Z_T$ given $\hat{\bar{\xi}}_B, \hat{\bar{\xi}}_T$ and $\theta$ is expressed as follows:

$$L(Z_B, Z_T|\hat{\bar{\xi}}_B, \hat{\bar{\xi}}_T, \theta) = \prod_{i=1}^{N} \prod_{j=1}^{n_B} P_{jzj}(\theta_i | \hat{\xi}_{Bij}) \prod_{k=1}^{n_T} P_{kz}(\theta_i | \hat{\xi}_{Tik})$$  \quad (8)

Using equations (4)–(6),

$$\hat{a}_{Tk} = \tilde{a}_{Tk}/h, \quad (9)$$

$$\hat{b}_{Tk} = h\tilde{b}_{Tk} + l, \quad (10)$$

and

$$\hat{\alpha}_{Tk} = \tilde{\alpha}_{Tk}/h, \quad (11)$$

we obtain the following likelihood

$$L(Z|h, l, \hat{\bar{\xi}}, \theta) = \prod_{i=1}^{N} \prod_{j=1}^{n_B} P_{jzj}(\theta_i | \hat{\xi}_{Bij}) \prod_{k=1}^{n_T} P_{kz}(\theta_i | \hat{\xi}_{Tik})$$  \quad (12)

where $Z = [Z_B|Z_T]$ and $\hat{\bar{\xi}} = [\hat{\bar{\xi}}_B|\hat{\bar{\xi}}_T]$. In Eq.(12) the unknown parameters are $h, l$ and $\theta$, and Noguchi (1986) estimated these variables by the joint maximum likelihood estimation method. Furthermore, Ogasawara (2001a) proposed a method maximizing the equation

$$L^*(h, l, \mu, \sigma^2|Z, \hat{\bar{\xi}}) \propto \prod_{i=1}^{N} \int_{-\infty}^{\infty} L_i(z_i|h, l, \hat{\bar{\xi}}, \theta_i) N(\theta_i|\mu, \sigma^2) \, d\theta_i$$  \quad (13)

to obtain $h, l, \mu$ and $\sigma^2$. Nuisance parameters $\theta$ are integrated over the prior distribution of $\theta$ in Eq.(13), where

$$L_i(z_i|h, l, \hat{\bar{\xi}}, \theta_i) = \prod_{j=1}^{n_B} P_{jzj}(\theta_i | \hat{\xi}_{Bij}) \prod_{k=1}^{n_T} P_{kz}(\theta_i | \hat{\xi}_{Tik})$$  \quad (14)

and $z_i = [z_{B1i} \cdots z_{Bni} z_{T1i} \cdots z_{Tni}]' = [z_{B1i} \cdots z_{Bni} z_{T1i} \cdots z_{Tni}]'$.

Although the likelihood was integrated over the prior distribution of $\theta_i$ by Ogasawara (2001a), $\theta_i$ can be marginalized by the posterior distribution of $\theta_i$ using the EM algorithm (Dempster, et al., 1977) to obtain the expected log-likelihood. The EM algorithm
is a numerical solution method which optimizes a function by repeating the E-step and M-step. Let \( h^{(t)} \) and \( l^{(t)} \) be the estimated \( h \) and \( l \) from the \( t \)-th M-step. As denoted by Tanner (1993) or Wang and Zeng (1998), the log of Eq.(12) integrated over the posterior distributions of nuisance parameter \( \theta_i \) can be expressed as follows:

\[
E_{\theta|h^{(t)},l^{(t)},\hat{\Xi},Z}[\ln L(h, l|\hat{\Xi}, Z, \theta)] = E_{\theta|h^{(t)},l^{(t)},\hat{\Xi},Z}[\ln L(Z|\hat{\Xi}, h, l, \theta)] \\
+ E_{\theta|h^{(t)},l^{(t)},\hat{\Xi},Z}[\ln L(Z|\Xi, \theta)] \\
- E_{\theta|h^{(t)},l^{(t)},\hat{\Xi},Z}[\ln L(Z|\hat{\Xi}, \theta)].
\]  

(15)

The first term of Eq.(15), the so-called Q function (Dempster, et al., 1977), is

\[
E_{\theta|h^{(t)},l^{(t)},\hat{\Xi},Z}[\ln L(Z|\hat{\Xi}, h, l, \Xi, \theta)] = \sum_{i=1}^{N} \int_{-\infty}^{\infty} \ln L_i(z_i|h_i,h,\hat{\Xi},\theta_i)\pi(\theta_i|h^{(t)},l^{(t)},\hat{\Xi},Z_i)\ d\theta_i,
\]

(16)

where \( \pi(\theta_i|h^{(t)},l^{(t)},\hat{\Xi},z_i) \) is the posterior distribution of \( \theta_i \) given \( h^{(t)}, l^{(t)}, \hat{\Xi} \) and \( z_i \). Also, the second term of Eq.(15) becomes as follows:

\[
E_{\theta|h^{(t)},l^{(t)},\hat{\Xi},Z}[\ln L(Z|\Xi, \theta)] = \ln L(h) + \ln L(l)
\]

(17)

under the assumption that \( h \) and \( l \) are independent of \( \theta \), since \( h \) and \( l \) can be regarded as item parameters in a broad sense, and item parameters are independent of person parameters. The third term of Eq.(15) becomes a constant.

Finally, at the \( t \)-th EM cycle, the posterior distribution of \( \theta_i (i=1, 2, \cdots, N) \) is first obtained by

\[
\pi(\theta_i|h^{(t)},l^{(t)},\hat{\Xi},Z_i) = \frac{L(z_i|h^{(t)},l^{(t)},\hat{\Xi},\theta_i)N(\theta_i|\hat{\theta}_{Bi},1)}{\int_{-\infty}^{\infty} L(z_i|h^{(t)},l^{(t)},\hat{\Xi},\theta_i)N(\theta_i|\hat{\theta}_{Bi},1)\ d\theta_i}
\]

(18)

in the E-step. The normal distribution with mean \( \hat{\theta}_{Bi} \) and variance 1 should be used for the prior distribution for \( \theta_i \), where \( \hat{\theta}_{Bi} \) is the examinee \( i \)'s latent trait estimated beforehand from the base test. In addition, instead of variance 1, the standard error of \( \hat{\theta}_{Bi} \) can be used for the variance of the prior distribution.

Next, in the M-step, the following function

\[
E_{\theta|h^{(t)},l^{(t)},\hat{\Xi},Z}[\ln L(h, l|\hat{\Xi}, Z, \theta)] \\
= \sum_{i=1}^{N} \int_{-\infty}^{\infty} \ln L_i(z_i|h_i,h,\hat{\Xi},\theta_i)\pi(\theta_i|h^{(t)},l^{(t)},\hat{\Xi},Z_i)\ d\theta_i \\
+ \ln L(h) + \ln L(l) + \text{const.}
\]

(19)

is maximized with respect to \( h \) and \( l \) to obtain \( h^{(t+1)} \) and \( l^{(t+1)} \). The integration over the posterior \( \theta \) distribution can be computed numerically using Gaussian quadrature points or equally spaced quadrature points. With \( Q \) quadrature points on the \( \theta \) scale, and with
no prior distribution for \( h \) and \( l \), Eq.(19) can be rewritten as

\[
f(h, l) = \sum_{i=1}^{N} \sum_{q=1}^{Q} \ln L_i(z_i|h, l, \hat{\Xi}, Y_q) W_q^{(l)},
\]

where \( Y_q(q = 1, 2, \cdots, Q) \) is a quadrature point and \( W_q^{(l)} = \pi(Y_q|h^{(l)}, l^{(l)}, \hat{\Xi}, z_i) \) \((q = 1, 2, \cdots, Q)\) is a corresponding weight.

In order to maximize Eq.(20), its partial derivatives with respect to the equating coefficients should be set to 0 and simultaneous equations should be solved to obtain the estimates of \( h \) and \( l \). The fundaments of the derivatives with respect to \( h \) and \( l \) of Eq.(20) are

\[
\frac{\partial \ln L_i(z_i|h, l, \hat{\Xi}, Y_q)}{\partial h} = \frac{\partial}{\partial h} \left\{ \sum_{j=1}^{n_B} \ln P_{jz_i}^2 (Y_q|\hat{\xi}_{Bj}) + \sum_{k=1}^{n_T} \ln P_{kz_i}^2 (Y_q|\hat{\xi}_{Tk}) \right\} \\
= \frac{\partial}{\partial h} \left\{ \text{const.} + \sum_{k=1}^{n_T} \ln P_{kz_i}^2 (Y_q|\hat{\xi}_{Tk}) \right\} \\
= \frac{n_T}{h^3} \left( Y_q - h\hat{\mu}_{Tk} - l - \frac{h\hat{z}_{ik}}{\hat{\sigma}_{Tk}} \right)
\]

and

\[
\frac{\partial \ln L_i(z_i|h, l, \hat{\Xi}, Y_q)}{\partial l} = \sum_{k=1}^{n_T} \frac{h^2}{h^2} \left( Y_q - h\hat{\mu}_{Tk} - l - \frac{h\hat{z}_{ik}}{\hat{\sigma}_{Tk}} \right),
\]

respectively. It is obvious from the expressions above that items of the base test do not contribute to maximization. The estimates of the base test items are used only when locating \( \theta_i(i = 1, 2, \cdots, N) \) in the E-step. To solve the simultaneous equations above set to 0, we can use the Newton–Raphson method to iteratively obtain the equating coefficients estimates for the current EM cycle. E-step and M-step above are repeated to convergence.

Although the standard errors for the estimates of equating coefficients would be of concern, as noted by Ogasawara (2001a), standard errors in the context of the EM algorithm have been discussed (Tanner, 1993, p.74). Therefore, we do not discuss them here, although several approaches are available (e.g. Baker, 1992; Jamshidian & Jennrich, 2000; Louis, 1972; Oakes, 1998). These methods considered standard errors within very simple models in which the number of parameters was less than ten. These methods seem to be unavailable for some complicated models, as mentioned by Baker (1992).

After estimating the equating coefficients, as described by Mislevy (1984, Equation 15), the trait distribution of all the common examinees can also be estimated iteratively as follows:

\[
p^{(r+1)}(\theta|h, l, \hat{\Xi}, Z) = \frac{1}{N} \sum_{i=1}^{N} \frac{L(z_i|h, l, \hat{\Xi}, \theta)p^{(r)}(\theta|h, l, \hat{\Xi}, Z)}{\int_{-\infty}^{\infty} L(z_i|h, l, \hat{\Xi}, \theta)p^{(r)}(\theta|h, l, \hat{\Xi}, Z) \, d\theta},
\]
although this is not a concern in our study.

If we let \( p^*(\theta | \hat{h}, \hat{l}, \Xi, Z) \) be the converged trait distribution, the expectation and variance of distribution can be obtained as follows:

\[
E[\theta | \hat{h}, \hat{l}, \Xi, Z] = \int_{-\infty}^{\infty} \theta p^*(\theta | \hat{h}, \hat{l}, \Xi, Z) \, d\theta,
\]

\[
V[\theta | \hat{h}, \hat{l}, \Xi, Z] = \int_{-\infty}^{\infty} \theta^2 p^*(\theta | \hat{h}, \hat{l}, \Xi, Z) \, d\theta - E[\theta | \hat{h}, \hat{l}, \Xi, Z]^2.
\]

2.3.2 Common Items Design

Numerous equating methods under the common items design have been discussed (e.g. Marco, 1977; Haebara, 1980; Stocking & Lord, 1983) in the dichotomous response case. In this study, we applied the Haebara (1980) method and Stocking & Lord (1983) method to the CRM. The key point of these two methods is to use the item response function (IRF) estimated from both the base and the target test. Haebara used the sum of the squared difference of two IRFs over the number of common items as the loss function and defined it multiplied by the weight and marginalized with respect to \( \theta \) as the weighted loss function. Finally, the equating coefficients which make the weighted loss function the minimum were solved numerically. In the Stocking and Lord (1983) method, the square of the sum of the difference of two IRFs was defined as the loss function.

The IRF (or item characteristic curve, ICC) is the probability of success of an item for a fixed trait \( \theta \) in the dichotomous IRT, although it can be interpreted as the conditional expectation of an item score when given a specific trait \( \theta \).

Although the IRF of the CRM can be expressed as

\[
E[Z_j|\theta] = \int_{-\infty}^{\infty} z_j P_{jz_j}(\theta | a_j, b_j, \alpha_j) \, dz_j,
\]

the variable \( z_j \) follows a normal distribution with mean \( \alpha_j(\theta - b_j) \) and variance \( \alpha_j^2 / a_j^2 \) by Eq.(1). Therefore, as stated by Wang and Zeng (1998), the expected item score under the CRM when \( \theta \) is given can be obtained as follows:

\[
E[Z_j|\theta] := T_j(\theta | a_j, b_j, \alpha_j) = \alpha_j(\theta - b_j),
\]

although it must be noted that parameter \( a_j \) does not appear in the equation above.

Let \( n \) be the number of common items, \( \lambda_{Bj} = [\hat{a}_{Bj}, \hat{b}_{Bj}, \hat{\alpha}_{Bj}]' \) represent the item parameter estimates for an item \( j \) from the base group data set, and \( \lambda_{Tj} = [\hat{a}_{Tj}, \hat{b}_{Tj}, \hat{\alpha}_{Tj}]' \) represent the item parameters estimated from the target group data set. Let \( \lambda_{\hat{T}j} = [\hat{a}_{\hat{T}j}, \hat{b}_{\hat{T}j}, \hat{\alpha}_{\hat{T}j}]' \) be \( \lambda_{Tj} (j = 1, 2, \ldots, n) \) equated onto the scale of the base group. Using equations (4)–(6), \( \lambda_{\hat{T}j} \) can be expressed as

\[
\hat{a}_{\hat{T}j} = \hat{a}_{Tj}/h,
\]

\[
\hat{b}_{\hat{T}j} = h \hat{b}_{Tj} + l,
\]

\[
\hat{\alpha}_{\hat{T}j} = \hat{\alpha}_{Tj}/h.
\]
The following equation
\[ q_j(h, l, \theta) = \frac{1}{2} \{ T_j(\theta | \hat{X}_B) - T_j(\theta | \hat{X}_T) \}^2 \tag{31} \]
can be used as the loss function of item \(j\), since \(T_j\)'s of the two groups are expected to be equal under the linear transformation of equations (28)-(30). Again, note that \(\hat{\alpha}_B\) and \(\hat{\alpha}_T\) are not used in Eq.(31). Then, the sum of Eq.(31) over the common items,
\[ q(h, l, \theta) = \sum_{j=1}^{n} q_j(h, l, \theta), \tag{32} \]
can be adopted as the loss function of all the common items. Finally, like Haebara (1980), we can define the following
\[ q^*(h, l) = \int_{-\infty}^{\infty} q(h, l, \theta)N(\theta|0, 1) \, d\theta \]
\[ = \int_{-\infty}^{\infty} \frac{1}{2} \sum_{j=1}^{n} \left\{ (\hat{\alpha}_B - \hat{\alpha}_T h) - \hat{\alpha}_T \frac{l}{h} \right\}^2 \frac{1}{\sqrt{2\pi}} \exp \left( -\frac{\theta^2}{2} \right) \, d\theta \]
\[ = \frac{1}{2} \sum_{j=1}^{n} \left\{ \left( \hat{\alpha}_B - \hat{\alpha}_T h \right)^2 + \left( \hat{\alpha}_B \hat{b}_B - \hat{\alpha}_T \hat{b}_T - \frac{\hat{\alpha}_T l}{h} \right)^2 \right\}, \tag{33} \]
as the weighted loss function. The form above happens to be very similar to the equating method proposed by Mayekawa and Suzuki (1988), who used the weighted sum of the squared difference of IRFs' logit transformations as the weighted loss function. Also, this expression above happens to be mathematically equivalent to the item logit method (Ogasawara, 2001b, Equation 17).

Differentiating Eq.(33) with respect to \(h\) and \(l\) and setting the results equal to 0, we obtain
\[ \frac{\partial q^*(h, l)}{\partial h} = \sum_{j=1}^{n} \left\{ \left( \hat{\alpha}_B - \hat{\alpha}_T \hat{b}_B \right) \frac{\hat{\alpha}_T h}{h^2} + \left( \hat{\alpha}_B \hat{b}_B - \hat{\alpha}_T \hat{b}_T \right) \frac{\hat{\alpha}_T l}{h^2} \right\} = 0, \tag{34} \]
and
\[ \frac{\partial q^*(h, l)}{\partial l} = -\sum_{j=1}^{n} \left\{ \left( \hat{\alpha}_B \hat{b}_B - \hat{\alpha}_T \hat{b}_T \right) \hat{\alpha}_T \frac{l}{h} \right\} = 0. \tag{35} \]

Solving the simultaneous equations above, we can obtain equating coefficients explicitly as follows:
\[ \hat{h} = \frac{\sum_{j=1}^{n} \hat{\alpha}_B}{\sum_{j=1}^{n} \hat{\alpha}_B \hat{\alpha}_T} \tag{36} \]
\[ \hat{l} = \frac{\sum_{j=1}^{n} (\hat{\alpha}_B \hat{b}_B - \hat{\alpha}_T \hat{b}_T) \hat{\alpha}_T}{\sum_{j=1}^{n} \hat{\alpha}_B \hat{\alpha}_T}. \tag{37} \]
In addition, if we use the following function

\[ q_T(h, l, \theta) = \frac{1}{2} \left[ \sum_{j=1}^{n} \left( T_j(\theta|\lambda_{B_j}) - T_j(\theta|\lambda_{T_j}^*) \right) \right]^2 \]  

(38)

as the loss function of all the common items, we obtain the weighted loss function as follows:

\[ q^-_{T}(h, l) = \int_{-\infty}^{\infty} q_T(h, l, \theta)N(\theta|0, 1) \, d\theta \]

\[ = \frac{1}{2} \left[ \left( \sum_{j=1}^{n} \left( \hat{\alpha}_{B_j} - \hat{\alpha}_{T_j} \right) \right)^2 + \sum_{j=1}^{n} \left( \hat{\alpha}_{B_j} \hat{b}_{B_j} - \hat{\alpha}_{T_j} \hat{b}_{T_j} - \frac{\hat{\lambda}_{T_j}}{h} \right) \right] \].  

(39)

The function above corresponds to the test characteristic curve method (Stocking & Lord, 1983) and is mathematically identical to the test logit method proposed by Ogasawara (2001b, Equation 26).

Differentiating Eq.(39) with respect to \( h \) and \( l \) and setting the results equal to zero, we can obtain equating coefficients as follows:

\[ \hat{h} = \frac{\sum_{j=1}^{n} \hat{\alpha}_{T_j}}{\sum_{j=1}^{n} \hat{\alpha}_{B_j}} \]  

(40)

\[ \hat{l} = \frac{\sum_{j=1}^{n} (\hat{\alpha}_{B_j} \hat{b}_{B_j} - \hat{\alpha}_{T_j} \hat{b}_{T_j})}{\sum_{j=1}^{n} \hat{\alpha}_{B_j}} \]  

(41)

The standard errors of the equating coefficient estimates can be obtained by using the method proposed by Ogasawara (2001b). He obtained the standard errors by the delta method using the asymptotic covariance matrix of the item parameter estimates. However, we do not discuss the standard errors here, since discussion on the asymptotic covariance matrix of the estimates in the context of the EM algorithm has not been finished yet, as mentioned in section 2.3.2.

3. SIMULATION STUDY

To confirm the accuracy of our methods, three sets of simulations were performed. One was under the common examinees design, and the other two were under the common items design.

3.1 Simulation under the Common Examinees Design

Using the notation defined in section 2.3.1, we prepared three different sample sizes (\( N=50, 100, 200 \)) of common examinees and three different numbers of items (\( n_B + n_T=20, 40, 80 \)), where the test length of the base test \( n_B \) was equal to that of the target test \( n_T \).

First, we generated item parameters \( \Xi_B, \Xi_T \). The natural logarithm of each discrimination parameter was generated from the normal distribution with a mean 0 and variance
0.09 \left( \ln \hat{\alpha}_{X_m} \sim N(0,0.09) \right) (X = B, T; m = j; k; j = 1, 2, \cdots, n_B; k = 1, 2, \cdots, n_T)). \] For difficulty parameters, each \( \hat{b}_{X_m} \) was drawn from \( N(0,1) \). Finally, the natural logarithm of \( \hat{\alpha}_{X_m} \) was generated from \( N(0,0.09) \) since \( \alpha \) was greater than 0. Then, we used an item response generation program, RESGEN (Muraki, 1992), to obtain continuous scores \( Z_B, Z_T \) under the condition that the latent trait distribution is standard normal. Finally, we estimated \( h \) and \( l \) given \( \Xi_B, \Xi_T, Z_B \) and \( Z_T \) using the method proposed in section 2.3.1. For each \( 3 \times 3 \) condition, we replicated the step 100 times. Most calibrations converged in less than five or six EM cycles.

### 3.2 Simulation under the Common Items Design

Using the new notation as represented in Figure 1, we prepared three different sample sizes \((N_B + N_T = 1000, 2000, 4000)\), the number of the base group examinees \( N_B \) being equal to that of the target group examinees \( N_T \). We also prepared three different test lengths \((n^* = n_1 + n_2 + n_3 = 15, 30, 60; n_1 = n_2 = n_3)\). Items in Test 2 are common items, and \( n_2 \) is the number of common items. Test 2 is often called an anchor test.

<table>
<thead>
<tr>
<th></th>
<th>Test 1</th>
<th>Test 2</th>
<th>Test 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Group</td>
<td>Item</td>
<td>Item</td>
<td>Item</td>
</tr>
<tr>
<td>Base</td>
<td>\vdots</td>
<td>( Z_{B1} )</td>
<td>( Z_{B2} )</td>
</tr>
<tr>
<td></td>
<td>( N_B )</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Target</td>
<td>\vdots</td>
<td>( Z_{T2} )</td>
<td>( Z_{T3} )</td>
</tr>
<tr>
<td></td>
<td>( N_T )</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 1: Notation For Section 3.2 And Section 4.2

First, we generated item parameters of Tests 1, 2 and 3. Each item parameter generation procedure was the same as the common examinees design simulation above. Next, using RESGEN (Muraki, 1992), we generated continuous scores for \( Z_{B1}, Z_{B2}, Z_{T2} \) and \( Z_{T3} \) (Figure 1) under the condition that the latent trait distribution is standard normal. Then, we estimated the item parameters of Test 1 and Test 2 from the base group data set \( Z_B = [Z_{B1}, Z_{B2}] \) using a procedure described by Wang and Zeng (1998), which estimates the item parameters of the CRM by the MML–EM method. Likewise, we estimated the item parameters of Test 2 and Test 3 from the target group data set \( Z_T = [Z_{T2}, Z_{T3}] \). Finally, we calculated \( h \) and \( l \) from two sets of item parameter estimates for Test 2 by applying two methods described in section 2.3.2. In addition, we replicated the step 100 times for each \( 3 \times 3 \) condition.

### 3.3 Simulation Results

In any of our simulation situations, true values of the estimated equating coefficients \( \hat{h} \)
and \( \hat{h} \) are 1 and 0, respectively. We adopted the mean difference (MD) and the root mean squared difference (RMSD) as indices for the accuracy of our methods. Smaller MD and RMSD values indicate better accuracy of the estimated results. The MD of \( \hat{h} \) and \( \hat{l} \) were defined as

\[
MD_h = \frac{\sum_{r=1}^{100} (\hat{h}_r - 1)}{100}
\]

and

\[
MD_l = \frac{\sum_{r=1}^{100} \hat{l}_r}{100},
\]

where \( \hat{h}_r \) and \( \hat{l}_r \) are the estimated values at the \( r \)th replication.

The RMSD for \( \hat{h} \) and \( \hat{l} \) were defined as follows:

\[
RMSD_h = \sqrt{\frac{\sum_{r=1}^{100} (\hat{h}_r - 1)^2}{100}}
\]

\[
RMSD_l = \sqrt{\frac{\sum_{r=1}^{100} \hat{l}_r^2}{100}}.
\]

Through simulation studies, corresponding means (\( M \)), standard deviations (\( SD \)), MDs, and RMSDs were calculated and are shown in Tables 1, 2 and 3, although the \( M \) of \( l \)'s is equal to the MD of \( l \)'s. Table 1 shows the results of the common examinees design, and Table 2 and Table 3 are those of the common items design: Table 2 shows the Haebara (1980) method, and Table 3 shows the Stocking and Lord (1983) method. In general, there was a tendency for the estimation accuracy to improve as the number of both observations and items increased, although there were some inharmonious cells in Tables 1, 2 and 3, due to sample variation. This fact supports the results of many previous studies.

When examining the accuracy of the equating coefficients in Table 1 individually, the \( h \) parameters were slightly overestimated. This indicates that our method proposed in section 2.3.1 has a slight bias, although it seems to be satisfactory in practice. We can see that biases dissolve as the number of items and observations increase. This fact implies that the proposed method has asymptotic unbiasedness. For \( l \) parameters, the results were mostly successful and the directions of their biases were random. The SDs and the RMSDs of \( h \) and \( l \) had a tendency to become smaller as the number of both observations and items increased. It seems that by doubling the number of items \( (n_B + n_T) \), the SDs of the equating coefficients are reduced by a factor of approximately \( 1/\sqrt{2} \), and the SD values when \( N = 200 \) are approximately half of those of \( N = 50 \).

In Table 2 and Table 3, the error for each cell includes estimation errors of the item parameters and errors from equating simultaneously because our simulation includes the true values recovery of the item parameters. Therefore, the error size in our simulation procedure is confounded with item parameter estimation errors. This means that the error size does not properly indicate the inaccuracy of the equating method proposed in section 2.3.2. However, it seems that the estimation results were accurate in general. The MD seems to approach \( 0 \) as the numbers of items and examinees increase. In addition, the
Table 1: Simulation Results Under Common Examinees Design

<table>
<thead>
<tr>
<th>$n_B + n_T$</th>
<th>$N$</th>
<th>$M$</th>
<th>$SD$</th>
<th>$MD$</th>
<th>$RMSD$</th>
<th>$h$</th>
<th>$SD$</th>
<th>$MD$</th>
<th>$RMSD$</th>
<th>$l$</th>
<th>$SD$</th>
<th>$MD$</th>
<th>$RMSD$</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>50</td>
<td>1.1013</td>
<td>0.0850</td>
<td>0.1103</td>
<td>0.1392</td>
<td>0.1069</td>
<td>0.0239</td>
<td>0.1096</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>100</td>
<td></td>
<td>1.0724</td>
<td>0.0629</td>
<td>0.0724</td>
<td>0.0959</td>
<td>0.0826</td>
<td>0.0367</td>
<td>0.0904</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>200</td>
<td></td>
<td>1.0400</td>
<td>0.0446</td>
<td>0.0400</td>
<td>0.0599</td>
<td>0.0641</td>
<td>-0.0288</td>
<td>0.0703</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>40</td>
<td>50</td>
<td>1.1063</td>
<td>0.0713</td>
<td>0.1063</td>
<td>0.1281</td>
<td>0.0993</td>
<td>0.0504</td>
<td>0.1114</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>100</td>
<td></td>
<td>1.0628</td>
<td>0.0552</td>
<td>0.0628</td>
<td>0.0836</td>
<td>0.0652</td>
<td>-0.0378</td>
<td>0.0754</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>200</td>
<td></td>
<td>1.0159</td>
<td>0.0351</td>
<td>0.0159</td>
<td>0.0385</td>
<td>0.0475</td>
<td>-0.0314</td>
<td>0.0569</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>80</td>
<td>50</td>
<td>1.1065</td>
<td>0.0576</td>
<td>0.1065</td>
<td>0.1211</td>
<td>0.0733</td>
<td>0.0520</td>
<td>0.0898</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>100</td>
<td></td>
<td>1.0414</td>
<td>0.0376</td>
<td>0.0414</td>
<td>0.0559</td>
<td>0.0553</td>
<td>0.0547</td>
<td>0.0778</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>200</td>
<td></td>
<td>1.0100</td>
<td>0.0278</td>
<td>0.0100</td>
<td>0.0295</td>
<td>0.0472</td>
<td>0.0484</td>
<td>0.0676</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 2: Simulation Results Under Common Items Design (Haebaer Method)

<table>
<thead>
<tr>
<th>$n^*$</th>
<th>$N_B + N_T$</th>
<th>$M$</th>
<th>$SD$</th>
<th>$MD$</th>
<th>$RMSD$</th>
<th>$h$</th>
<th>$SD$</th>
<th>$MD$</th>
<th>$RMSD$</th>
<th>$l$</th>
<th>$SD$</th>
<th>$MD$</th>
<th>$RMSD$</th>
</tr>
</thead>
<tbody>
<tr>
<td>15</td>
<td>1000</td>
<td>0.9930</td>
<td>0.0533</td>
<td>-0.0070</td>
<td>0.0538</td>
<td>0.0517</td>
<td>0.0034</td>
<td>0.0518</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>2000</td>
<td>0.9829</td>
<td>0.0328</td>
<td>-0.0171</td>
<td>0.0370</td>
<td>0.0401</td>
<td>0.0137</td>
<td>0.0424</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>4000</td>
<td>1.0037</td>
<td>0.0251</td>
<td>0.0037</td>
<td>0.0254</td>
<td>0.0267</td>
<td>-0.0027</td>
<td>0.0268</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>30</td>
<td>1000</td>
<td>0.9933</td>
<td>0.0345</td>
<td>-0.0107</td>
<td>0.0361</td>
<td>0.0477</td>
<td>-0.0115</td>
<td>0.0491</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>2000</td>
<td>0.9948</td>
<td>0.0272</td>
<td>-0.0052</td>
<td>0.0277</td>
<td>0.0303</td>
<td>-0.0023</td>
<td>0.0304</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>4000</td>
<td>0.9955</td>
<td>0.0156</td>
<td>-0.0045</td>
<td>0.0162</td>
<td>0.0259</td>
<td>0.0067</td>
<td>0.0268</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>60</td>
<td>1000</td>
<td>0.9948</td>
<td>0.0298</td>
<td>-0.0052</td>
<td>0.0303</td>
<td>0.0418</td>
<td>-0.0191</td>
<td>0.0460</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>2000</td>
<td>1.0034</td>
<td>0.0203</td>
<td>0.0034</td>
<td>0.0206</td>
<td>0.0276</td>
<td>-0.0066</td>
<td>0.0282</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>4000</td>
<td>0.9964</td>
<td>0.0120</td>
<td>-0.0036</td>
<td>0.0125</td>
<td>0.0204</td>
<td>-0.0004</td>
<td>0.0225</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 3: Simulation Results Under Common Items Design (Stocking & Lord Method)

<table>
<thead>
<tr>
<th>$n^*$</th>
<th>$N_B + N_T$</th>
<th>$M$</th>
<th>$SD$</th>
<th>$MD$</th>
<th>$RMSD$</th>
<th>$h$</th>
<th>$SD$</th>
<th>$MD$</th>
<th>$RMSD$</th>
<th>$l$</th>
<th>$SD$</th>
<th>$MD$</th>
<th>$RMSD$</th>
</tr>
</thead>
<tbody>
<tr>
<td>15</td>
<td>1000</td>
<td>0.9942</td>
<td>0.0511</td>
<td>-0.0058</td>
<td>0.0515</td>
<td>0.0397</td>
<td>0.0042</td>
<td>0.0399</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>2000</td>
<td>1.0050</td>
<td>0.0396</td>
<td>0.0050</td>
<td>0.0400</td>
<td>0.0456</td>
<td>-0.0010</td>
<td>0.0456</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>4000</td>
<td>0.9979</td>
<td>0.0294</td>
<td>-0.0021</td>
<td>0.0294</td>
<td>0.0263</td>
<td>0.0049</td>
<td>0.0268</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>30</td>
<td>1000</td>
<td>1.0001</td>
<td>0.0461</td>
<td>0.0001</td>
<td>0.0461</td>
<td>0.0542</td>
<td>-0.0007</td>
<td>0.0542</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>2000</td>
<td>1.0034</td>
<td>0.0358</td>
<td>0.0034</td>
<td>0.0360</td>
<td>0.0350</td>
<td>0.0026</td>
<td>0.0351</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>4000</td>
<td>0.9953</td>
<td>0.0231</td>
<td>-0.0047</td>
<td>0.0235</td>
<td>0.0266</td>
<td>-0.0028</td>
<td>0.0267</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>60</td>
<td>1000</td>
<td>0.9999</td>
<td>0.0419</td>
<td>-0.0001</td>
<td>0.0419</td>
<td>0.0462</td>
<td>0.0001</td>
<td>0.0462</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>2000</td>
<td>1.0015</td>
<td>0.0298</td>
<td>0.0015</td>
<td>0.0299</td>
<td>0.0329</td>
<td>-0.0012</td>
<td>0.0329</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>4000</td>
<td>1.0011</td>
<td>0.0209</td>
<td>0.0011</td>
<td>0.0209</td>
<td>0.0245</td>
<td>0.0014</td>
<td>0.0245</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

$SD$ and $RMSD$ become smaller with larger item and observation numbers. This suggests that the item parameter recovery becomes more successful as the numbers of items and examinees increase. This fact leads to equating accuracy.

As for the difference between the two methods, the Stocking & Lord method had high accuracy for $h$. On the contrary, the result of the Haebaer method for $l$ was better than that of the Stocking & Lord method.
4. NUMERICAL EXAMPLES BY REAL DATA

We analyzed an actual item response data set\(^1\) whose test length was 30 and observation number was 1280. The latent trait measured by the test was "vocational competency". All the items had seven categories and were scored on a scale of 0 to 6. Here, we recode 0 and 6 as 0.01 and 5.99, because \( z_{ij} = \ln(x_{ij}/(k_j - x_{ij})) \) transformation does not allow \( x_{ij} \) to be both 0 and \( k_j \). Samejima (1973) called this situation the "open response situation". In contrast, if a subject is allowed to mark the end points, she called the situation the "closed response situation". We confirmed the numerical performances of equating coefficient estimates from the real data set in the following two situations.

4.1 Real Data Analysis Under Common Examinees Design

Using the notation as represented in Figure 2,

1. We regarded the odd number items and even number items of the test as Base Test and Target Test, respectively.
2. Sampling 50 observations, we defined them as Group 2 (common examinees).
3. Randomly dividing the remaining 1230 observations into two groups, we labeled one of them as Group 1 and the other as Group 3.
4. We estimated the item parameters of Base Test and Target Test from the data set \( Z_B = [Z_{1B}|Z_{2B}] \) and \( Z_T = [Z_{2T}|Z_{3T}] \), respectively.
5. Using the common examinees response matrix \( Z_2 = [Z_{2B}|Z_{2T}] \), we estimated the equating coefficients by the method proposed in section 2.3.1.

We processed Group 1 examinees’ responses for Test 2 and Group 3 examinees’ responses for Test 1 as missing data. The procedure set above was repeated 100 times.

<table>
<thead>
<tr>
<th>Group</th>
<th>Test Obs._Item</th>
<th>Base 1_15</th>
<th>Target 1_15</th>
</tr>
</thead>
<tbody>
<tr>
<td>Group 1</td>
<td>615</td>
<td>( Z_{1B} )</td>
<td></td>
</tr>
<tr>
<td>Group 2</td>
<td>50</td>
<td>( Z_{2B} )</td>
<td>( Z_{2T} )</td>
</tr>
<tr>
<td>Group 3</td>
<td>615</td>
<td>( Z_{3T} )</td>
<td></td>
</tr>
</tbody>
</table>

Figure 2: Notation For Section 4.1

\(^1\) This data was provided by JMA Management Center Inc. Assessment Research Center. The author thanks them for their support.
4.2 Real Data Analysis Under Common Items Design

Using the notation already denoted in Figure 1,
1. Randomly dividing the 1280 observations into 2 groups, one of them was regarded as the Base Group and the other as the Target Group.
2. Selecting 10 items randomly, we regarded them as Test 2 (anchor test).
3. We randomly allotted a half of the remaining 20 items to Test 1 and the other half to Test 3.
4. We estimated the item parameters of Test 1 and Test 2 from the data set $Z_B = [Z_{B1} | Z_{B2}]$. We also estimated the item parameters of Test 2 and Test 3 from $Z_T = [Z_{T2} | Z_{T3}]$.
5. We estimated the equating coefficients from two sets of item parameter estimates for Test 2 using the two methods proposed in section 2.3.2.

Base Group subjects’ responses for Test 3 and Target Group subjects’ responses for Test 1 were ignored in the procedure above. We repeated this 100 times to obtain 100 pairs of equating coefficient estimates.

4.3 Real Data Analysis Results

Corresponding $M$s, $SD$s, $MD$s and $RMSD$s of estimation results under both common examiners and items designs are shown in Table 4. Figure 3 shows the scatter plot for 100 pairs of equating coefficient estimates of the three proposed methods under two equating designs.

The true values of estimated $h$ and $l$ are 1 and 0, respectively. In Table 4, each $MD$ seemed to be close to 0. $MD$ of $h$ under the common examiners design was slightly overestimated, conforming with simulation studies. In addition, $SD$s and $RMSD$s values were small enough for practical use. Figure 3 suggests that the estimated $h$ and $l$ seem to be normally distributed with a moderate negative correlation under each design. Moreover, the estimates of the two methods under the common items design were more centered than those under the common examiners design, although they cannot be compared directly. Finally, the Haebaara method yielded a better result for $l$, and the Stocking & Lord method was more accurate for $h$, which is in agreement with simulation studies.

5. CONCLUSION

This study made linking tests in which items were calibrated by the CRM possible under both the common examiners and common items designs by estimating the equating coefficients. The proposed methods seemed to be accurate in practice. Although the method under the common examiners design which applies the MML-EM does not require highly sophisticated programming skills, its repeating optimization procedure seems to be comparatively complex for the average test practitioners to use. However, the common items design is the popular choice in almost all linking situations. The two proposed methods under the common items design are very easy to use for typical test practitioners
Table 4: Results of Real Data Analysis

<table>
<thead>
<tr>
<th>Design</th>
<th>$M$</th>
<th>$SD$</th>
<th>$MD$</th>
<th>$RMSD$</th>
<th>$SD$</th>
<th>$MD$</th>
<th>$RMSD$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Common Examinees</td>
<td>1.0351</td>
<td>0.0962</td>
<td>0.0351</td>
<td>0.1024</td>
<td>0.1016</td>
<td>-0.0504</td>
<td>0.1179</td>
</tr>
<tr>
<td>Haabera Method</td>
<td>1.0075</td>
<td>0.0380</td>
<td>0.0075</td>
<td>0.0388</td>
<td>0.0349</td>
<td>0.0072</td>
<td>0.0357</td>
</tr>
<tr>
<td>Stocking &amp; Lord Method</td>
<td>1.0061</td>
<td>0.0376</td>
<td>0.0061</td>
<td>0.0371</td>
<td>0.0361</td>
<td>0.0096</td>
<td>0.0373</td>
</tr>
</tbody>
</table>

Figure 3: Scatter Plot Of Estimated Equating Coefficients
(The symbol ■ represents the estimates under the common examinees design. The symbols × and ◊ represent the estimates by the Haabera method and Stocking & Lord method, respectively.)

since the IRF of the CRM has a very simple form, enabling estimates to be obtained without numerical iterations.

Acknowledgements

The author is grateful to two reviewers for their useful comments and suggestions on earlier versions of this paper.
REFERENCES


(Received April 2002, Revised August 2002)