Economic Review (Otaru University
of Commerce), Vol.51, No.1, 1-23,
July, 2000.

Asymptotic Standard Errors of IRT Equating Coefficients Using Moments

Haruhiko Ogasawara

Abstract

The asymptotic standard errors of the IRT equating coefficients given by the mean/sigma, mean/mean and mean/geometric mean methods are derived when the two-parameter logistic model holds and item parameters are obtained by the marginal maximum likelihood estimation. The case of two nonequivalent examinee-groups and the case of single group are considered. The numerical examples show that the mean/mean and mean/geometric mean methods are superior to the mean/sigma method. The results also show that the number of quadrature points in the numerical approximation to the integration of ability parameters is crucial to the estimation of the asymptotic standard errors.

Keywords: Equating, IRT, asymptotic standard errors, mean/sigma method, mean/mean method, marginal maximum likelihood estimation.

The author is indebted to Michael J. Kolen who has made available the real data analyzed in this article. Request for reprints should be sent to Haruhiko Ogasawara, Otaru University of Commerce, 3-5-21, Midori, Otaru 047-8501 Japan. Email: hogasa@res.otaru-uc.ac.jp

Test equating becomes necessary when the scale of a test is to be compared with the scale of another test. If the two tests are composed of items whose characteristics are given by applying item response theory (IRT) separately to each test, the results of one of the tests can not directly be compared to those of another test. This comes from the fact that usually the abilities of examinees in a test are standardized with mean zero and unit variance to remove the indeterminacy of an IRT model. Hence, the estimates of item parameters (and abilities if necessary) in one test should be transformed to the scales of the parameters of another test by equating.

In IRT equating, the method of common items is often used in the above situation. The common items may be part of each test (internal items) or external ones. A simple equating procedure in such situations with common items is that of using moments (means and standard deviations) of the estimates of the common items. From the property of the IRT model, the transformation for the parameters in equating should be linear. The equating coefficients are estimated as the slope and intercept in the linear transformation. Marco (1977) used the means and standard deviations of difficulty parameters. This is called the mean/sigma (*m/s*) method. Shiba (1978) used the means of discrimination parameters in addition to those of difficulty parameters. Loyd and Hoover (1980) used a similar method in the Rasch model. This is called the mean/mean (*m/m*) method. As a variation of the *m/m* method, Mislevy and Bock (1990) (see also Kolen & Brennan, 1995, Ch.6) proposed a method using the geometric means of discrimination parameters and the arithmetic means of difficulty parameters. This is called the mean/geometric mean (*m/gm*) method in this article.

Other methods using the item/test characteristic curves (Haebara, 1980; Stocking & Lord, 1983) have also been developed. These methods are more sophisticated than the methods using moments of the item parameters. However, the methods by moments are easy and simple to apply in practice and seem to give similar results to those by item characteristic methods in some situations (see e.g., Baker & Al-Karni, 1991; Hattori, 1998).

The purpose of this article is to obtain the asymptotic standard errors of the estimates of the equating coefficients by the *m/s*, *m/m* and

m/gm methods with the assumption of the two-parameter logistic model and to compare their results with each other. Baker and Al-Karni (1991) indicated that the *m/m* method is more stable than the *m/s* method. This will be made clear using the asymptotic behavior of the estimated equating coefficients in simulated data. The results by IRT in samples depend on the estimation methods of item (and ability) parameters. Consequently, the asymptotic standard errors of the estimates of equating coefficients also depend on the asymptotic variances and covariances of the estimates of the IRT parameters. Historically, the joint maximum likelihood estimation of item and ability parameters was first developed (see e.g., Lord & Novick, 1968, Ch.17). The asymptotic variances of the estimates of item parameters by this method are usually estimated from the information matrix of the estimated item parameters assuming that ability parameters are given (see e.g., Lord, 1980, p.191; Wainer & Thissen, 1982). Since they are underestimates of exact asymptotic variances, the standard error of an equating in IRT using the underestimates of the asymptotic variances-covariances is also an underestimate (see e.g. Lord, 1982). The exact asymptotic variances by the joint maximum likelihood estimation may be obtained by assuming that both of the numbers of items and examinees become large. However, this is an unrealistic assumption.

In this article, we deal with the case when item parameters are estimated by the marginal maximum likelihood (Bock & Lieberman, 1970; Bock & Aitkin, 1981) in which abilities are integrated out from the model. Thus, the standard asymptotic theory applies to the estimates of the item parameters given by the method. In the following sections, we will consider the case of internal common items. The application to the external common items is straightforward and will be discussed in the final section.

Equating Methods Using Moments

We deal with the case of two independent nonequivalent examinee-groups (Groups 1 and 2): the examinees in Group 1 take Test 1 and those in Group 2 take Test 2. The results for single examinee-group are essentially the same as far as the results in this

section are concerned. Assume that the probability of a correct/incorrect response to the j-th common item by the i-th examinee in Group 1 is described by the two-parameter logistic model:

$$P_1(x_{1ij}|\theta_{1i},a_{1j},b_{1j})=\frac{\exp\{-Da_{1j}(\theta_{1i}-b_{1j})(1-x_{1ij})\}}{1+\exp\{-Da_{1j}(\theta_{1i}-b_{1j})\}}, \quad (1)$$

$(i=1,\ldots,N_1; j=1,\ldots,p)$,

where $x_{1ij}=1$ denotes a correct response and $x_{1ij}=0$ an incorrect response in the above situation; θ_{1i} is the ability score for the i-th examinee in Group 1; N_1 is the number of examinees in Group 1; a_{1j} and b_{1j} are the discrimination and difficulty parameters, respectively, for the j-th common item of Test 1; p is the number of common items; D=1.7 is a costant. For Group 2, we have

$$P_2(x_{2ij}|\theta_{2i},a_{2j},b_{2j})=\frac{\exp\{-Da_{2j}(\theta_{2i}-b_{2j})(1-x_{2ij})\}}{1+\exp\{-Da_{2j}(\theta_{2i}-b_{2j})\}}, \quad (2)$$

$(i=1,\ldots,N_2; j=1,\ldots,p)$,

where notations are similarly defined. The true values of the parameters in the j-th common item in Group 1, a_{1j} and b_{1j}, are the same as those in Group 2, a_{2j} and b_{2j}, respectively, if they are appropriately transformed. In addition to the p common items, we assume that there are q_1-p and q_2-p unique items in Tests 1 and 2, respectively. That is, Tests 1 and 2 consist of q_1 and q_2 items, respectively. The parameters for the unique items are a_{1j} and b_{1j}, $j=p+1,\ldots,q_1$, for Test 1 and a_{2j} and b_{2j}, $j=p+1,\ldots,q_2$, for Test 2.

Equating is supposed to be performed such that the scale in Test 2 is transformed to that in Test 1. For model identification, we assume $\theta_{1i} \overset{\text{i.i.d.}}{\sim} N(0,1), i=1,..,N_1$ and $\theta_{2i} \overset{\text{i.i.d.}}{\sim} N(0,1), i=1,..,N_2$. Let

$$\theta_{2i}^{*}=A\theta_{2i}+B, \quad a_{2j}^{*}=a_{2j}/A \text{ and } b_{2j}^{*}=Ab_{2j}+B. \quad (3)$$

Then, from (2)

$$P_2(x_{2ij}|\theta_{2i},a_{2j},b_{2j}) = P_2(x_{2ij}|\theta^*_{2i},a^*_{2j},b^*_{2j}),$$
$$(i=1,...,N_2; j=1,...,q_2). \tag{4}$$

For the p common items, if A and B are appropriately chosen,

$$a_{1j} = a^*_{2j} \text{ and } b_{1j} = b^*_{2j}, \ (j=1,..,p). \tag{5}$$

However, the equations of (5) hold only in populations. In samples, the relationships in (5) are at most approximate ones. Therefore, the task is to estimate A and B such that (5) holds as closely as possible. The estimates of A and B using moments are defined as follows:

$$\hat{A}_s = \sqrt{\frac{\sum_{j=1}^{p}\hat{b}^2_{1j} - (1/p)(\sum_{j=1}^{p}\hat{b}_{1j})^2}{\sum_{j=1}^{p}\hat{b}^2_{2j} - (1/p)(\sum_{j=1}^{p}\hat{b}_{2j})^2}}, \tag{6}$$

$$\hat{B}_s = (1/p)\sum_{j=1}^{p}\hat{b}_{1j} - \hat{A}_s(1/p)\sum_{j=1}^{p}\hat{b}_{2j}$$

for the *m/s* method,

$$\hat{A}_m = \sum_{j=1}^{p}\hat{a}_{2j} / \sum_{j=1}^{p}\hat{a}_{1j}, \tag{7}$$

$$\hat{B}_m = (1/p)\sum_{j=1}^{p}\hat{b}_{1j} - \hat{A}_m(1/p)\sum_{j=1}^{p}\hat{b}_{2j}$$

for the *m/m* method and

$$\hat{A}_g = (\prod_{j=1}^{p}\hat{a}_{2j} / \hat{a}_{1j})^{1/p}, \tag{8}$$

$$\hat{B}_g = (1/p)\sum_{j=1}^{p}\hat{b}_{1j} - \hat{A}_g(1/p)\sum_{j=1}^{p}\hat{b}_{2j}$$

for the *m/gm* method. Note that in populations $A_s = A_m = A_g$ and $B_s = B_m = B_g$.

Asymptotic Standard Errors of Equating Coefficients

From the definitions of the equating coefficients, we see that they

are functions of the item parameters. Thus, the asymptotic variances-covariances of the estimates of the coefficients are obtained from the asymptotic variance-covariance matrix of the estimates of the item parameters by using the delta method. Let

$\underline{\alpha}_{1j}=(a_{1j},b_{1j})'$, $(j=1,\ldots,q_1)$, $\underline{\alpha}_1=(\underline{\alpha}'_{11},\ldots,\underline{\alpha}'_{1q_1})'$,

$\underline{\alpha}_{2j}=(a_{2j},b_{2j})'$, $(j=1,\ldots,q_2)$, $\underline{\alpha}_2=(\underline{\alpha}'_{21},\ldots,\underline{\alpha}'_{2q_2})'$ and

$\underline{\alpha}=(\underline{\alpha}_1',\underline{\alpha}_2')'$.

(Note that $\underline{\alpha}$ represents the whole item parameters including the parameters for unique items in Tests 1 and 2.) Then, the asymptotic variance-covariance matrix for the vector of the estimates $(\hat{A}*,\hat{B}*)'$ is

$$\mathrm{acov}(\hat{A}*,\hat{B}*)'=\frac{\partial(A*,B*)'}{\partial\underline{\alpha}'}\mathrm{acov}(\underline{\hat{\alpha}})\frac{\partial(A*,B*)}{\partial\underline{\alpha}}, \qquad (9)$$

where $A*$ and $B*$ denote a pair of the equating coefficients. Because Group 1 is assumed to be independent of Group 2 in the case of two nonequivalent groups,

$$\mathrm{acov}(\underline{\hat{\alpha}})=\begin{bmatrix}\mathrm{acov}(\underline{\hat{\alpha}}_1) & O\\ O & \mathrm{acov}(\underline{\hat{\alpha}}_2)\end{bmatrix}. \qquad (10)$$

In the case of single group, since the same examinees take two tests, we have

$$\mathrm{acov}(\underline{\hat{\alpha}})=\begin{bmatrix}\mathrm{acov}(\underline{\hat{\alpha}}_1) & \mathrm{acov}(\underline{\hat{\alpha}}_1;\underline{\hat{\alpha}}_2)\\ \mathrm{acov}(\underline{\hat{\alpha}}_2;\underline{\hat{\alpha}}_1) & \mathrm{acov}(\underline{\hat{\alpha}}_2)\end{bmatrix}, \qquad (11)$$

where $\mathrm{acov}(\underline{\hat{\alpha}}_2;\underline{\hat{\alpha}}_1)$ is the covariance matrix of $\underline{\hat{\alpha}}_2$ with respect to $\underline{\hat{\alpha}}_1$ and $\mathrm{acov}(\underline{\hat{\alpha}}_1;\underline{\hat{\alpha}}_2)=\{\mathrm{acov}(\underline{\hat{\alpha}}_2;\underline{\hat{\alpha}}_1)\}'$.

The partial derivatives in (9) are obtained by elementary calculus and will be provided in Appendix for completeness. Notice that the partial derivatives in (9) with respect to the parameters in the unique items in Tests 1 and 2 are zero since $A*$ and $B*$ do not include them. The estimate of (9) is given by substituting the estimates of the item

parameters for the true values in the right-hand side of (9)

The remaining task is to obtain $\mathrm{acov}(\hat{\underline{\alpha}})$ in (10) and (11). First, we investigate the case of two nonequivalent groups. The estimates of the item parameters for the q_1 items including unique ones in Test 1 are obtained by maximizing the following marginal likelihood with the assumption of multivariate normality for abilities:

$$L_1^* = \prod_{i=1}^{N_1} \int_{-\infty}^{+\infty} L_{1i}(\underline{\alpha}_1 | \theta_{1i}, \underline{x}_{1i}) h(\theta_{1i}) d\theta_{1i} \quad (12)$$

where

$$L_{1i}(\underline{\alpha}_1 | \theta_{1i}, \underline{x}_{1i}) = \prod_{j=1}^{q_1} P_1(x_{1ij} | \theta_{1i}, \underline{\alpha}_{1j}) \quad (13)$$

with $\underline{x}_{1i} = (x_{1i1}, \ldots, x_{1iq_1})'$ and

$$h(\theta_{1i}) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{\theta_{1i}^2}{2}). \quad (14)$$

Since the integration in (12) is difficult to obtain, it is approximated by a numerical one to any desired accuracy as follows:

$$L_1^* \cong L_1 = \prod_{i=1}^{N_1} \sum_{m=1}^{r} L_{1i}(\underline{\alpha}_1 | y_m, \underline{x}_{1i}) H(y_m) \equiv \prod_{i=1}^{N_1} f_1(\underline{x}_{1i} | \underline{\alpha}_1), \quad (15)$$

where $y_1, \ldots, y_r$ are the quadrature points in the range for θ_{1i} and $H(y_m)$ are the weights for the quadrature points, which are proportional to $h(y_m)$ with $\sum_{m=1}^{r} H(y_m) = 1$ and an adjustment to satisfy $\sum_{m=1}^{r} y_m^2 H(y_m) = 1$. Let $l_1 = \ln L_1$. Then, the maximization of L_1 in (15) is given by solving the equations:

$$\frac{\partial l_1}{\partial \underline{\alpha}_{1j}} = \sum_{i=1}^{N_1}\sum_{m=1}^{r} \frac{\partial \ln P_1(x_{1ij}|y_m, \underline{\alpha}_{1j})}{\partial \underline{\alpha}_{1j}} \times \frac{L_{1i}(\underline{\alpha}_1|y_m, \underline{x}_{1i})H(y_m)}{f_1(\underline{x}_{1i}|\underline{\alpha}_1)}$$

$$= \sum_{i=1}^{N_1}\sum_{m=1}^{r} \{x_{1ij} - P_1(x_{1ij}=1|y_m, \underline{\alpha}_{1j})\} D \begin{pmatrix} y_m - b_{1j} \\ -a_{1j} \end{pmatrix} \phi_1(y_m|\underline{x}_{1i}, \underline{\alpha}_1) \quad (16)$$

$$\equiv \sum_{i=1}^{N_1} \underline{g}_{1ij} = \underline{0}, \quad (j=1,\ldots,q_1),$$

where

$$\phi_1(y_m|\underline{x}_{1i}, \underline{\alpha}_1) \equiv \frac{L_{1i}(\underline{\alpha}_1|y_m, \underline{x}_{1i})H(y_m)}{f_1(\underline{x}_{1i}|\underline{\alpha}_1)}, \quad (17)$$

$$(m=1,\ldots,r; i=1,\ldots,N_1)$$

is the posterior probability of y_m given $\underline{\alpha}_1$ and $\underline{x}_{1i}$. Similar results are obtained for Group 2 with Test 2 using similar notations:

$$\frac{\partial l_2}{\partial \underline{\alpha}_{2j}} = \sum_{i=1}^{N_2}\sum_{m=1}^{r} \{x_{2ij} - P_2(x_{2ij}=1|y_m, \underline{\alpha}_{2j})\}$$

$$\times D \begin{pmatrix} y_m - b_{2j} \\ -a_{2j} \end{pmatrix} \phi_2(y_m|\underline{x}_{2i}, \underline{\alpha}_2) \equiv \sum_{i=1}^{N_2} \underline{g}_{2ij}, \quad (18)$$

$(j=1,\ldots,q_2)$.

The asymptotic variance-covariance matrix for $\hat{\underline{\alpha}}_1$ is obtained from the inverse of the information matrix for the item parameters. However, since 2^{q_1} patterns in $\underline{x}_{1i}$ are required to derive the exact information matrix (see, Bock & Lieberman, 1970), only the observed patterns for $\underline{x}_{1i}$ are used as an approximation to the exact one, that is,

$$\hat{I}(\hat{\underline{\alpha}}_1) = \sum_{i=1}^{N_1} \underline{g}_{1i} \underline{g}_{1i}'|_{\underline{\alpha}_1 = \hat{\underline{\alpha}}_1} \quad (19)$$

where

$$\underline{g}_{1i} = (\underline{g}_{1i1}', \ldots, \underline{g}_{1iq_1}')' \quad (20)$$

(see (16)). The estimate of the asymptotic variance-covariance matrix for $\hat{\underline{\alpha}}_1$ is obtained as:

$$\hat{\text{acov}}(\hat{\underline{\alpha}}_1) = (\hat{I}(\hat{\underline{\alpha}}_1))^{-1}. \tag{21}$$

Similarly, we have

$$\hat{\text{acov}}(\hat{\underline{\alpha}}_2) = (\hat{I}(\hat{\underline{\alpha}}_2))^{-1}. \tag{22}$$

Note that (21) and (22) hold also in the case of single group with some adaptations such as $N = N_1 = N_2$ ($\underline{\alpha}_1$ and $\underline{\alpha}_2$ are assumed to be estimated separately in each test even in the case of single group).

Finally, we derive $\text{acov}(\hat{\underline{\alpha}}_2; \hat{\underline{\alpha}}_1)$ in (11) which is required in the case of single group. By using the Taylor expansions of the observed gradient vectors of (16) and (18) at the true values of the parameters with large $N(= N_1 = N_2)$, we have approximately

$$\hat{\underline{\alpha}}_1 - \underline{\alpha}_1 \cong (I(\hat{\underline{\alpha}}_1))^{-1} \sum_{i=1}^{N} \underline{g}_{1i}, \; \hat{\underline{\alpha}}_2 - \underline{\alpha}_2 \cong (I(\hat{\underline{\alpha}}_2))^{-1} \sum_{i=1}^{N} \underline{g}_{2i}. \tag{23}$$

Hence, taking the expectation of $(\hat{\underline{\alpha}}_2 - \underline{\alpha}_2)(\hat{\underline{\alpha}}_1 - \underline{\alpha}_1)'$ in large samples and noting that $\underline{g}_{1i}$ and $\underline{g}_{2j} (i \neq j)$ are independent, we have

$$\text{acov}(\hat{\underline{\alpha}}_2; \hat{\underline{\alpha}}_1) = (I(\hat{\underline{\alpha}}_2))^{-1} \text{E}(\sum_{i=1}^{N} \underline{g}_{2i} \underline{g}_{1i}')(I(\hat{\underline{\alpha}}_1))^{-1}. \tag{24}$$

For the estimates of $I(\hat{\underline{\alpha}}_1)$ and $I(\hat{\underline{\alpha}}_2)$ in (24), we can again use (21) and (22).

The term of $\text{E}(\cdot)$ in the right-hand side of (24) is obtained as:

$$\text{E}(\sum_{i=1}^{N} \underline{g}_{2i} \underline{g}_{1i}') = N \sum_{k_1=1}^{2^{q_1}} \sum_{k_2=1}^{2^{q_2}} \frac{f_{12}(\underline{x}_{k_1}, \underline{x}_{k_2} | \underline{\alpha})}{f_1(\underline{x}_{k_1} | \underline{\alpha}_1) f_2(\underline{x}_{k_2} | \underline{\alpha}_2)} \times \frac{\partial f_2(\underline{x}_{k_2} | \underline{\alpha}_2)}{\partial \underline{\alpha}_2} \frac{\partial f_1(\underline{x}_{k_1} | \underline{\alpha}_1)}{\partial \underline{\alpha}_1'} \tag{25}$$

where

$$f_{12}(\underline{x}_{k_1}, \underline{x}_{k_2} | \underline{\alpha}) = \sum_{m=1}^{r} L_{1i}(\underline{\alpha}_1 | y_m, \underline{x}_{k_1}) L_{2i}(\underline{\alpha}_2 | y_m, \underline{x}_{k_2}) H(y_m); \quad (26)$$

$\underline{x}_{k_1}$ is the k_1-th possible response pattern for $\underline{x}_{1i}$, $(k_1 = 1, \ldots, 2^{q_1})$; $\underline{x}_{k_2}$ is the k_2-th possible response pattern for $\underline{x}_{2i}$, $(k_2 = 1, \ldots, 2^{q_2})$. The values 2^{q_1} and 2^{q_2} become soon large with moderate q_1 and q_2. Therefore, a practical estimate of (25) is simply $\sum_{i=1}^{N} \underline{g}_{2i} \underline{g}_{1i}'$ with $\underline{\alpha} = \hat{\underline{\alpha}}$, which is an approximation using only the observed patterns of $\underline{x}_{1i}$ and $\underline{x}_{2i}$, $(i = 1, \ldots, N)$.

Mean Standard Error of Equated Scores

The ability score in Group 2, θ_{2i}, is transformed to the score in Group 1 by using the estimates of the equating coefficients $\hat{A}_*$ and $\hat{B}_*$:

$$\hat{\theta}_{2i}^{*} = \hat{A}_* \theta_{2i} + \hat{B}_*, \quad (i = 1, \ldots, N_2). \quad (27)$$

To evaluate the overall stability of $\hat{A}_*$ and $\hat{B}_*$, it is convenient to calculate the standard error of the equated score at θ_{2i}:

$$\begin{aligned} \mathrm{SE}(\hat{\theta}_{2i}^{*}) &= \sqrt{\mathrm{avar}(\hat{A}_* \theta_{2i} + \hat{B}_*)} \\ &= \sqrt{\mathrm{avar}(\hat{A}_*)\theta_{2i}^2 + 2\,\mathrm{acov}(\hat{A}_*; \hat{B}_*)\theta_{2i} + \mathrm{avar}(\hat{B}_*)}. \end{aligned} \quad (28)$$

The mean standard error of the equated score is obtained from the integration over the distribution of θ_{2i}:

$$\begin{aligned} &\sqrt{\int_{-\infty}^{+\infty} \mathrm{avar}(\hat{A}_* \theta_{2i} + \hat{B}_*) h(\theta_{2i}) d\theta_{2i}} \\ &\qquad = \sqrt{\mathrm{avar}(\hat{A}_*) + \mathrm{avar}(\hat{B}_*)} \end{aligned} \quad (29)$$

The estimate of (29) is given by replacing the true values of the parameters in (29) by their estimates (see also, Kolen & Brennan, 1995, Ch.7).

Numerical Examples

To confirm the accuracy of the estimated standard errors for the equating coefficients, we have performed a simulation with true values. The first half of the simulation is for the case of two nonequivalent groups and the second half for the case of single group. In the first half, the numbers of common items are set at 10 or 15 with the same numbers of unique items in Tests 1 and 2. That is, Tests 1 and 2 have 20 or 30 items including the common items. The population values of discrimination parameters were randomly generated by the uniform distribution with the range (.3, 1.3). The population difficulty parameters were also randomly generated by the normal distribution $N(0, 1)$. The observed values of item responses in Group 1 were generated by using the probability function of (1) with the random number following $N(0, 1)$ for θ_{1i}. For Group 2, $\theta_{2i} \overset{\text{i.i.d.}}{\sim} N(.5, 1.2^2)$ was employed for the generation of the observed responses. Consequently, if estimation is exact, $\hat{A}_* = 1.2$, $\hat{B}_* = .5$ should be obtained. The number of examinees in each group is 1,000 (Case A) or 2,000 (Case B) when the number of common items is 10; and 1,000 when the number of common items is 15 (Case C). When the number of common items is 10, the same set of population values are used for the cases of N=1,000 and N=2,000. The numbers of quadrature points in the numerical approximation of the integration of ability parameters are 5, 10 or 15. The estimation of the equating coefficients was repeated 100 times in each condition. That is, 100 estimates for each coefficient were obtained with 100 estimates of its asymptotic standard error.

Table 1. Means of estimated equating coefficients for nonequivalent groups; number of sets of samples =100, population values for A (B)=1.2(.5).

	Case A			Case B			Case C		
Number of common items:	10			10			15		
Number of observations:	1,000			2,000			1,000		
Number of quadrature points:									
	5	10	15	5	10	15	5	10	15
A_s	1.189	1.214	1.210	1.175	1.203	1.199	1.118	1.179	1.185
B_s	.522	.505	.502	.518	.503	.500	.502	.496	.504
A_m	1.168	1.205	1.203	1.168	1.207	1.205	1.118	1.189	1.192
B_m	.505	.499	.497	.513	.507	.506	.503	.499	.507
A_g	1.172	1.206	1.204	1.170	1.206	1.205	1.119	1.189	1.192
B_g	.508	.500	.498	.515	.507	.505	.503	.499	.507

Tables 1 through 5 show the results for two nonequivalent groups. Table 1 shows the means of the estimated coefficients over 100 sets of samples. The table indicates that the estimates are somewhat biased when the number of quadrature points is 5. By increasing the number as large as 10, the biases are to a large extent reduced. The results of N=2,000 (Case B) are not so different from those of N=1,000 (Case A). Table 2 shows the results of theoretical and simulated standard errors for Case A. The *SD* is the standard deviation of the estimates of a coefficient or a statistic (the mean standard error of equated scores) over 100 sets of samples. The *M* of *SE* is the mean of estimated standard errors over 100 sets of samples. The *SD* of *SE* is the standard deviation of the estimated standard errors. If the estimated asymptotic standard errors are close to exact values, the *M*'s of *SE* should be close to the corresponding *SD*'s which are the actual standard deviation of the estimates and the *SD*'s of *SE* should be small. From the table, we see that when the number of quadrature points is 5, the asymptotic standard errors for $\hat{B}_*$ seem to be underestimates. However, they become rather accurate when the number of quadrature points is as large as 10. Among the three methods, *m/s*, *m/m* and *m/gm*, the *m/s* method is always inferior to the

other two methods. This is clearly shown in the large standard errors for $\hat{A}_S$, which supports the discussion of Baker and Al-Karni (1991). Table 3 shows the results for Case B, which are similar to Table 2 except the overall level of values. Note that the standard errors are proportional to $1/\sqrt{N}$. Thus, we see that the values of *SD* and *M* of *SE* in Table 3 are approximately $1/\sqrt{2}$ of corresponding values in Table 2 (notice that the same population values for item parameters are used in Cases A and B).

Table 2. Results for nonequivalent groups (Case A); number of common items = 10, number of observations in each sample = 1,000, number of sets of samples = 100.

Number of quadrature points	5			10			15		
	SD	*M* of *SE*	*SD* of *SE*	*SD*	*M* of *SE*	*SD* of *SE*	*SD*	*M* of *SE*	*SD* of *SE*
(1) Equating coefficients									
A_s	.122	.124	.018	.128	.129	.019	.129	.130	.019
B_s	.099	.079	.010	.084	.086	.009	.084	.086	.009
A_m	.053	.050	.003	.062	.061	.003	.062	.062	.003
B_m	.088	.066	.006	.080	.076	.005	.080	.076	.005
A_g	.054	.054	.003	.063	.063	.004	.063	.064	.004
B_g	.084	.060	.004	.073	.071	.003	.073	.070	.003
(2) Mean of standard error of equated scores									
m/s	.157	.147	.020	.153	.155	.020	.154	.156	.020
m/m	.103	.083	.005	.101	.097	.006	.101	.098	.006
m/gm	.100	.080	.005	.096	.095	.005	.097	.095	.005

Note. *SD* = standard deviation of estimates of a parameter or a statistic; *M* of *SE* = mean of estimated standard errors; *SD* of *SE* = standard deviation of estimated standard errors.

Table 3. Results for nonequivalent groups (Case B); number of common items = 10, number of observations in each sample = 2,000, number of sets of samples = 100.

Number of quadrature points	5			10			15		
	SD	*M* of *SE*	*SD* of *SE*	*SD*	*M* of *SE*	*SD* of *SE*	*SD*	*M* of *SE*	*SD* of *SE*
(1) Equating coefficients									
A_s	.081	.084	.008	.087	.087	.008	.088	.088	.009
B_s	.073	.054	.005	.056	.059	.004	.055	.059	.004
A_m	.035	.035	.001	.041	.042	.002	.042	.043	.002
B_m	.067	.046	.002	.050	.053	.002	.050	.053	.002
A_g	.039	.038	.002	.045	.044	.002	.046	.045	.002
B_g	.065	.041	.002	.046	.049	.002	.046	.049	.002
(2) Mean of standard error of equated scores									
m/s	.109	.099	.009	.104	.106	.009	.104	.106	.009
m/m	.075	.058	.002	.065	.068	.002	.065	.069	.003
m/gm	.076	.056	.002	.065	.066	.002	.065	.067	.002

Note. *SD* = standard deviation of estimates of a parameter or a statistic; *M* of *SE* = mean of estimated standard errors; *SD* of *SE* = standard deviation of estimated standard errors.

Table 4. Results for nonequivalent groups (Case C); number of common items = 15, number of observations in each sample = 1,000, number of sets of samples = 100.

Number of quadrature points	5			10			15		
	SD	*M* of *SE*	*SD* of *SE*	*SD*	*M* of *SE*	*SD* of *SE*	*SD*	*M* of *SE*	*SD* of *SE*
(1) Equating coefficients									
A_s	.051	.047	.005	.059	.056	.005	.060	.060	.006
B_s	.094	.049	.005	.072	.063	.004	.074	.065	.004
A_m	.046	.040	.002	.056	.052	.003	.056	.055	.003
B_m	.092	.044	.003	.069	.060	.002	.071	.062	.002
A_g	.045	.039	.002	.054	.050	.002	.054	.054	.003
B_g	.092	.044	.003	.069	.060	.002	.071	.062	.002
(2) Mean of standard error of equated scores									
m/s	.107	.068	.006	.093	.084	.006	.095	.088	.006
m/m	.103	.060	.003	.089	.079	.003	.091	.082	.003
m/gm	.102	.059	.003	.088	.078	.003	.089	.082	.003

Note. *SD* = standard deviation of estimates of a parameter or a statistic; *M* of *SE* = mean of estimated standard errors; *SD* of *SE* = standard deviation of estimated standard errors.

Table 5. Correlations between estimated equating coefficients (Case A); number of common items = 10, number of observations in each sample = 1,000, number of sets of samples = 100, number of quadrature points = 10.

	A_s	B_s	A_m	B_m	A_g	B_g
A_s	1.00	.68 (.03)	.38 (.03)	-.32 (.09)	.58 (.03)	-.20 (.08)
B_s	.64	1.00	.17 (.03)	.33 (.12)	.30 (.04)	.45 (.10)
A_m	.41	.18	1.00	.31 (.04)	.94 (.01)	.32 (.04)
B_m	-.36	.34	.29	1.00	.16 (.07)	.98 (.003)
A_g	.59	.27	.95	.12	1.00	.23 (.06)
B_g	-.26	.44	.30	.98	.18	1.00

Note. The lower half indicates the correlations of estimates of the coefficients. The upper half indicates the means (standard deviations) of the estimated asymptotic correlations for the estimates of the coefficients.

Table 4 gives the results for Case C, where the number of common items is 15. Surprisingly, the differences between the three methods which were observed in Tables 2 and 3 have almost disappeared, though the *m/s* method is still the worst one. Note that the tendency of the underestimates of $\hat{B}_*$ is stronger than those in Table 2 and 3 when the number of quadrature points is 5. Table 5 gives the observed correlations of the estimates of the coefficients, and the means (standard deviations) of the estimated asymptotic correlations over 100 sets of samples. The actual correlations are close to mean theoretical values. The pairs of ($\hat{A}_g$ and $\hat{A}_m$) and ($\hat{B}_g$ and $\hat{B}_m$) have high correlations within each pair, which suggests the closeness of the *m/m* and *m/gm* methods.

Tables 6 and 7 show the results for single group (Case A'). The population values for item parameters are the same as those for Cases A and B. The number of observations is 1,000. Since the same examinees respond to the items in Tests 1 and 2, θ_{1i} is set equal to θ_{2i} when random responses are generated. Thus, if the estimation is exact, $\hat{A}_* = 1$ and $\hat{B}_* = 0$ should be obtained. In Tables 6 and 7, we observe the similar tendencies which were shown in Tables 1 and 2. However, the standard errors for Case A' are reduced from those for Case A. This is theoretically expected from the signs of partial derivatives (see Appendix) and the non-negligible positive covariances between $\underline{\hat{\alpha}}_1$ and $\underline{\hat{\alpha}}_2$ for the case of single group.

Table 6. Means of estimated equating coefficients for single group (Case A'); number of common items = 10, number of observations in each sample = 1,000, number of samples = 100, population values for A (B) = 1 (0).

	Number of quadrature points		
	5	10	15
A_s	1.013	1.007	1.007
B_s	-.011	-.004	-.003
A_m	1.010	1.000	1.000
B_m	-.012	-.005	-.005
A_g	1.009	1.000	.999
B_g	-.012	-.006	-.005

Table 7. Results for single group (Case A'); number of common items = 10, number of observations in each sample = 1,000, number of samples = 100.

Number of quadrature points	5			10			15		
	SD	*M* of *SE*	*SD* of *SE*	*SD*	*M* of *SE*	*SD* of *SE*	*SD*	*M* of *SE*	*SD* of *SE*
(1) Equating coefficients									
A_s	.092	.101	.015	.091	.100	.014	.091	.101	.014
B_s	.061	.044	.004	.037	.039	.003	.037	.038	.003
A_m	.036	.038	.002	.038	.038	.002	.038	.038	.002
B_m	.066	.056	.006	.049	.051	.006	.049	.051	.006
A_g	.040	.042	.002	.042	.042	.003	.042	.042	.003
B_g	.062	.051	.005	.044	.046	.005	.044	.046	.005
(2) Mean of standard error of equated scores									
m/s	.110	.110	.014	.098	.107	.014	.098	.108	.014
m/m	.075	.067	.006	.062	.064	.005	.062	.064	.005
m/gm	.074	.066	.005	.061	.063	.005	.061	.063	.005

Note. *SD* = standard deviation of estimates of a parameter or a statistic.; *M* of *SE* = mean of estimated standard errors; *SD* of *SE* = standard deviation of estimated standard errors.

Table 8. Results for Kolen and Brennan's (1995) data; number of common items = 12, number of items in each test = 36, numbers of observations in each group = 1655 (Test X) and 1638 (Test Y), number of quadrature points = 10.

	Estimates	*SE*	Asymptotic correlations					
A_s	1.009	.070	1.00					
B_s	-.375	.069	-.06	1.00				
A_m	.961	.041	.57	-.30	1.00			
B_m	-.349	.078	.26	.92	-.28	1.00		
A_g	.970	.044	.73	-.20	.94	-.10	1.00	
B_g	-.354	.075	.21	.94	-.29	.995	-.14	1.00

Mean standard error of equated scores

m/s: .099, *m/m*: .088, *m/gm*: .087

Note. *SE* = standard error of estimates.

Table 8 shows the results for a real data set. The data from Kolen and Brennan (1995, Appendix B) are used: Tests X and Y consisting of 36 items in each test have 12 internal common items and were administered to 1,655 and 1,638 examinees, respectively. The equating was performed by assuming that the groups are independent nonequivalent ones. The transformation in the equating was from the scale of Test X to that of Test Y in the two-parameter logistic model. Ten quadrature points were used for the numerical integration of abilities. The estimated coefficients for the *m/s* method are somewhat different from those for the *m/m* and *m/gm* methods. To the contrary of the simulated results, the standard error for $\hat{B}_s$ is smaller than those for $\hat{B}_m$ and $\hat{B}_g$. However, the mean standard error of equated scores for the *m/s* method is greater than those for the *m/m* and *m/gm* methods as was the case for simulated data. The estimated asymptotic correlations show a close relationship between *m/m* and *m/gm* methods.

Conclusion

The simulated results in the previous section are based on

restricted conditions. However, the results are rather clear and indicates that when the number of common items are small such as 10, the results of *m/s* method are inferior to those by the *m/m* and *m/gm* methods. The differences between three methods seem to decrease with the increase of the number of common items. Except for the unusual case when only the estimates of difficulties are available, we have no reason to employ the *m/s* method. The *m/m* method is recommended from its simplicity among the three methods as long as the evidence of the superiority of the *m/gm* method is not provided.

The marginal likelihood estimation of item parameters employs numerical integration. The estimates of the equating coefficients are directly influenced by the number of quadrature points in the numerical integration. The number should be as large as 10.

Discussion

Up to now, the situation of internal common items has been assumed. If external common items are used, the asymptotic covariance matrix of (10) and (11) should be reformulated in the following way. We assume the same number of common item as before. That is , the p common items are supposed to constitute the anchor test (Test 3). Tests 1 and 2 are composed of only unique items whose numbers are $q_1 - p$ and $q_2 - p$, respectively. The difference between this situation and that of internal common items is that the estimation of the item parameters are performed separately for Tests 1, 2 and 3 in the case of external common items. The parameters of the common items may be estimated jointly with those for Test 1 or Test 2. For this case, the situation becomes essentially equivalent to that with internal common items as long as the asymptotic behavior of the estimates of equating coefficients are concerned.

Let $\underline{\alpha}_1$ and $\underline{\alpha}_2$ be the vectors of the item parameters for Group 1 (Tests 1 and 3) and Group 2 (Tests 2 and 3), respectively as was the case for internal common items. The subvectors in $\underline{\alpha}_1$ and $\underline{\alpha}_2$ are defined:

$\underline{\alpha}_1 = (\underline{\beta}_1', \underline{\gamma}_1')'$ and $\underline{\alpha}_2 = (\underline{\beta}_2', \underline{\gamma}_2')'$, where $\underline{\beta}_1 = (a_{11}, b_{11}, \cdots, a_{1p}, b_{1p})'$ and $\underline{\beta}_2 = (a_{21}, b_{21}, \cdots, a_{2p}, b_{2p})'$ are the parameters for Test 3 (the anchor test), while $\underline{\gamma}_1 = (a_{1,p+1}, b_{1,p+1}, \cdots, a_{1q_1}, b_{1q_1})'$ and $\underline{\gamma}_2 = (a_{2,p+1}, b_{2,p+1}, \cdots, a_{2q_2}, b_{2q_2})'$ are the parameters for Tests 1 and 2, respectively. Let $\underline{\alpha} = (\underline{\alpha}_1', \underline{\alpha}_2')'$ as before. Then, the asymptotic variance-covariance matrix of $\hat{\underline{\alpha}}$ for the case of two nonequivalent groups becomes

$$\mathrm{acov}(\hat{\underline{\alpha}}) = \begin{bmatrix} \mathrm{acov}(\hat{\underline{\alpha}}_1) & O \\ O & \mathrm{acov}(\hat{\underline{\alpha}}_2) \end{bmatrix} \tag{30}$$

where

$$\mathrm{acov}(\hat{\underline{\alpha}}_k) = \begin{bmatrix} (I(\hat{\underline{\beta}}_k))^{-1} & (I(\hat{\underline{\beta}}_k))^{-1}\mathrm{E}(\sum_{i=1}^{N_k} \underline{g}_{\beta ki}\underline{g}_{\gamma ki}') \times (I(\hat{\underline{\gamma}}_k))^{-1} \\ (I(\hat{\underline{\gamma}}_k))^{-1}\mathrm{E}(\sum_{i=1}^{N_k} \underline{g}_{\gamma ki}\underline{g}_{\beta ki}') \times (I(\hat{\underline{\beta}}_k))^{-1} & (I(\hat{\underline{\gamma}}_k))^{-1} \end{bmatrix}, \tag{31}$$

$(k = 1, 2)$

with $\underline{g}_{\beta ki}$ and $\underline{g}_{\gamma ki}$ being the subvectors in $\underline{g}_{ki}$ (see (20) with (16) and (18)) for the parameters $\underline{\beta}_k$ and $\underline{\gamma}_k$, respectively. In the case of single group, the asymptotic cross covariance matrix for $\hat{\underline{\alpha}}_2$ with respect to $\hat{\underline{\alpha}}_1$ becomes

$$\text{acov}(\underline{\hat{\alpha}}_2;\underline{\hat{\alpha}}_1) =$$

$$\begin{bmatrix} (I(\underline{\hat{\beta}}_2))^{-1}\mathrm{E}(\sum_{i=1}^{N}\underline{g}_{\beta 2i}\underline{g}_{\beta 1i}') \times (I(\underline{\hat{\beta}}_1))^{-1}, & (I(\underline{\hat{\beta}}_2))^{-1}\mathrm{E}(\sum_{i=1}^{N}\underline{g}_{\beta 2i}\underline{g}_{\gamma 1i}') \times (I(\underline{\hat{\gamma}}_1))^{-1} \\ (I(\underline{\hat{\gamma}}_2))^{-1}\mathrm{E}(\sum_{i=1}^{N}\underline{g}_{\gamma 2i}\underline{g}_{\beta 1i}') \times (I(\underline{\hat{\beta}}_1))^{-1}, & (I(\underline{\hat{\gamma}}_2))^{-1}\mathrm{E}(\sum_{i=1}^{N}\underline{g}_{\gamma 2i}\underline{g}_{\gamma 1i}') \times (I(\underline{\hat{\gamma}}_1))^{-1} \end{bmatrix} \quad (32)$$

The estimates of (31) and (32) are given by substituting the estimates of the parameters for their true values, and the observed values for $\mathrm{E}(\cdot)$. Since the partial derivatives of the equating coefficients with respect to $\underline{\gamma}_1$ and $\underline{\gamma}_2$ are zero, only the upper-left submatrices in (31) and (32) are used in actual computation for $\text{avâr}(\hat{A}_*)$ and $\text{avâr}(\hat{B}_*)$. However, other submatrices become necessary when we consider the asymptotic variances and covariances of equated item parameters and their functions in Tests 1 and 2.

Appendix The Partial Derivatives of the Equating Coefficients with respect to the Item Parameters

For the *m/s* method (see (6)), the nonzero partial derivatives are

$$\frac{\partial A_s}{\partial b_{1j}} = \frac{A_s(b_{1j} - (1/p)\sum_{k=1}^{p} b_{1k})}{\sum_{k=1}^{p} b_{1k}^2 - (1/p)(\sum_{k=1}^{p} b_{1k})^2},$$

$$\frac{\partial A_s}{\partial b_{2j}} = \frac{-A_s(b_{2j} - (1/p)\sum_{k=1}^{p} b_{2k})}{\sum_{k=1}^{p} b_{2k}^2 - (1/p)(\sum_{k=1}^{p} b_{2k})^2},$$

$$\frac{\partial B_s}{\partial b_{1j}} = \frac{1}{p} - \frac{\partial A_s}{\partial b_{1j}} \times \frac{\sum_{k=1}^{p} b_{2k}}{p}, \quad \text{(A1)}$$

$$\frac{\partial B_s}{\partial b_{2j}} = -\frac{\partial A_s}{\partial b_{2j}} \times \frac{\sum_{k=1}^{p} b_{2k}}{p} - \frac{A_s}{p}, \quad (j = 1,\ldots,p).$$

For the *m/m* method (see (7)), the nonzero partial derivatives are

$$\frac{\partial A_m}{\partial a_{1j}} = -\frac{\sum_{k=1}^{p} a_{2k}}{(\sum_{k=1}^{p} a_{1k})^2}, \quad \frac{\partial A_m}{\partial a_{2j}} = \frac{1}{\sum_{k=1}^{p} a_{1k}},$$

$$\frac{\partial B_m}{\partial a_{1j}} = -\frac{\partial A_m}{\partial a_{1j}} \times \frac{\sum_{k=1}^{p} b_{2k}}{p}, \quad \frac{\partial B_m}{\partial a_{2j}} = -\frac{\partial A_m}{\partial a_{2j}} \times \frac{\sum_{k=1}^{p} b_{2k}}{p},$$

$$\frac{\partial B_m}{\partial b_{1j}} = \frac{1}{p}, \quad \frac{\partial B_m}{\partial b_{2j}} = -\frac{A_m}{p}, \quad (j = 1,\ldots,p). \qquad \text{(A2)}$$

For the *m/gm* method (see (8)), the nonzero partial derivatives are

$$\frac{\partial A_g}{\partial a_{1j}} = -\frac{A_g}{p a_{1j}}, \quad \frac{\partial A_g}{\partial a_{2j}} = \frac{A_g}{p a_{2j}}, \quad \frac{\partial B_g}{\partial a_{1j}} = -\frac{\partial A_g}{\partial a_{1j}} \times \frac{\sum_{k=1}^{p} b_{2k}}{p},$$

$$\frac{\partial B_g}{\partial a_{2j}} = -\frac{\partial A_g}{\partial a_{2j}} \times \frac{\sum_{k=1}^{p} b_{2k}}{p}, \quad \frac{\partial B_g}{\partial b_{1j}} = \frac{1}{p},$$

$$\frac{\partial B_g}{\partial b_{2j}} = -\frac{A_g}{p}, \quad (j = 1,\ldots,p). \qquad \text{(A3)}$$

References

Baker, F. B., & Al-Karni, A. (1991). A comparison of two procedures for computing IRT equating coefficients. *Journal of Educational Measurement, 28*, 147-162.

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., & Lieberman, M. (1970). Fitting a response model for *n* dichotomously scored items. *Psychometrika, 35*, 179-197.

Haebara, T. (1980). Equating logistic ability scales by a weighted least squares method. *Japanese Psychological Research, 22*(3), 144-149.

Hattori, T. (1998). Equating the parameters for the generalized partial credit model – Minimum χ^2 and item characteristic curve methods – . *Proceeding of the 62nd Annual Meeting of the Japanese Psychological Association*, 417. (in Japanese)

Kolen, M. J., & Brennan, R. L. (1995). *Test equating: Methods and practices*. New-York: Springer.

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

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

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

Lord, F. M. (1982). Standard error of an equating by item response theory., *Applied Psychological Measurement, 6*, 463-472.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison Wesley.

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

Shiba, S. (1978). Construction of a scale for acquisition of word meaning. *Bulletin of Faculty of Education, University of Tokyo, 17*, 47-58. (in Japanese)

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

Wainer, H., & Thissen, D. (1982). Some standard errors in item response theory, *Psychometrika, 47*, 397-412.