ED 463 333 TM 033 765

AUTHOR	Kim, Dong-In; Brennan, Robert; Kolen, Michael
TITLE	A Comparison of IRT Equating and Beta 4 Equating.
PUB DATE	2002-04-00
NOTE	27p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (New Orleans, LA, April 2-4, 2002). Some figures may not reproduce well.
PUB TYPE	Numerical/Quantitative Data (110) -- Reports - Evaluative (142) -- Speeches/Meeting Papers (150)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	Comparative Analysis; Criteria; Elementary Education; *Elementary School Students; *Equated Scores; *Item Response Theory; True Scores
IDENTIFIERS	*Equipercentile Equating; Iowa Tests of Basic Skills

ABSTRACT

Four equating methods were compared using four equating criteria: first-order equity (FOE), second-order equity (SOE), conditional mean squared error (CMSE) difference, and the equipercentile equating property. The four methods were: (1) three parameter logistic (3PL) model true score equating; (2) 3PL observed score equating; (3) beta 4 true score equating; and (4) beta 4 observed score equating. Four subtests of the Iowa Tests of Basic Skills, with two forms of each test and a random sample of 3,000 examinees for each form were used. Results show that true score equating more closely achieved FOE than observed score equating, while observed score equating more closely achieved SOE, CMSE difference, and equipercentile equating property than true score equating. Among the four equating methods, 3PL observed most closely achieved SOE and had the smallest CMSE difference, and beta 4 observed was the method that most closely met the equipercentile equating property. (Contains 6 tables, 7 figures, and 26 references.) (SLD)

A Comparison of IRT Equating and Beta 4 Equating[1]

Dong-In Kim
CTB/McGraw-Hill

Robert Brennan
Michael Kolen
University of Iowa

[1] Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, April 2002

Four equating methods, 3PL true score equating, 3PL observed score equating, beta 4 true score equating, and beta 4 observed score equating were compared using four equating criteria, first-order equity (FOE), second-order equity (SOE), conditional-mean-squared-error (CMSSE) difference, and the equipercentile equating property. True score equating more closely achieved FOE than observed score equating, while Observed score equating more closely achieved SOE, CMSE difference, and equipercentile equating property than true score equating. Among the four equating methods, 3PLO most closely achieved SOE and had the smallest CMSE difference, and B4O was the method that most closely met the equipercentile equating property.

A Comparison of IRT Equating and Beta 4 Equating

Many testing programs apply unidimensional Item Response Theory (IRT) models to assemble tests and equate test forms (Kolen & Brennan, 1995). Although the use of IRT equating requires strong assumptions made by IRT models, if these assumptions hold, IRT equating has many advantages over conventional equating procedures. Cook and Eignor (1991) described several theoretical and practical advantages of IRT equating. For example, the group invariance property is guaranteed in IRT true score equating if the IRT model holds.

A strong true score model, called the four-parameter beta compound binomial model, or beta 4, was developed by Lord (1965). Let

$$f(x) = \int h(x \mid \tau) g(\tau) \; d\tau \,, \qquad (1)$$

where x designates the observed score and τ designates the true score. This beta 4 model assumes that the distribution of true scores $g(\tau)$ is the four-parameter beta distribution, and the conditional distribution of the observed scores given true score $h(x \mid \tau)$ is the compound binomial distribution. Then, the observed score distribution $f(x)$ based on the above two assumptions is the four-parameter beta compound binomial distribution. Beta 4 equating, based on the above true score model, has two potential advantages compared to IRT equating. First, the assumptions of beta 4 are weaker than the assumptions of IRT. IRT is based on two strong assumptions, unidimensionality and local independence, which are often not met with real data (Lord, 1980). Second, compared to IRT equating, beta 4 equating requires estimating a much small number of parameters, at most five.

To date, there has been no comparison study between beta 4 equating and IRT equating. The comparison of two true score equating methods based on different psychometric models, beta 4 and IRT, can be helpful to more clearly understand the characteristics of true score equating. Similarly a comparison between the observed score equating methods based on beta 4 and IRT can be helpful to better understand the characteristics of observed score equating.

In many equating studies, traditional equating methods, such as equipercentile equating and linear equating, have been compared with IRT true score equating under various conditions (Cook, Eignor, & Schmitt, 1988; Eignor, Stocking, & Cook, 1986, 1995; Harris & Kolen, 1986; Kolen & Harris, 1990; Lawrence & Dorans, 1990; Livingston, Dorans, & Wright, 1990; Skaggs, 1990; Skaggs & Lissitz, 1986). But, IRT true score equating and IRT observed score equating have been compared in only a few studies (Kolen, 1981; Lord & Wingersky; 1984; Han, Kolen, & Pohlmann, 1997). Furthermore, while Kolen (1981), and Han, Kolen, & Pohlmann (1997) found some differences between IRT true and observed score equating methods, Lord and Wingersky (1984) reported similar results. In short, it is not clear whether true score equating results are likely to be different from observed score equating results. Also, there has been no comparison study between true score equating and observed score equating based on the beta 4 procedure. This study is intended to clarify the similarities and differences between true score equating and observed score equating.

Psychometric Models

The strong true score model (Lord, 1965) models the observed score distribution under the assumptions about the true score distribution and distribution of measurement error given true score for a population of examinees. This model is appropriate for a test, which consists of multiple-choice items and is scored number-correct. With K dichotomously scored test items, beta 4 can be expressed as follows:

$$f(x) = \int_{l}^{u} h(x \mid \tau, K) g(\tau \mid \alpha, \beta, l, u) \; d\tau, \tag{2}$$

where $g(\tau \mid \alpha, \beta, l, u)$ denotes the four-parameter beta distribution, and $h(x \mid \tau, K)$ represents the compound binomial distribution, τ is the proportion-correct true score, α and β are shape parameters, and l and u are lower limit and upper limit parameters. The four-parameter beta distribution (true score distribution) is a generalization of the family of beta distributions. When l =0 and u=1, the four-parameter beta distribution is the beta distribution. To simplify computations for the compound binomial distribution (error score distribution), a two-term approximation to the compound binomial distribution was suggested by Lord (1965).

Item Response Theory (IRT) supposes that an examinee's performance can be predicted by his or her latent trait, or ability (Lord, 1980). This performance is not predicted at the test score level, but at the item level. The relationship between the examinee's observed response pattern and the examinee's latent trait is described by a mathematical function (i.e., item response function, trace line, or item characteristic curve). This relationship is established based on two strong assumptions (Lord, 1980): unidimensionality and local independence. Two types of parameters determine the probability of success on item i, $p_i(\theta)$: one is the ability parameter, and the other is item parameters. The general and commonly used item response model for dichotomous items is Birnbaum's three-parameter logistic model (3PL) (Lord & Novick, 1968):

$$p_i(\theta) = c_i + (1 - c_i) \frac{1}{1 + e^{-1.7 a_i (\theta - b_i)}}, \tag{3}$$

where θ is ability parameter, a_i is item discrimination parameter, b_i is item difficulty parameter, and c_i is item guessing parameter. The number 1.7 is a scaling constant imposed to make the logistic function close to a normal ogive function.

Equating Methods

3PL True Score Equating (3PLT)

When Form X and Form Y are both measures of the same ability θ, their number-correct true scores are associated with the same θ through their test characteristic functions (Lord, 1980):

$$\tau_X(\theta_j) = \sum_{i=1}^{K} p_{ij}(\theta_j \mid a_i, b_i, c_i), \;\; \tau_Y(\theta_j) = \sum_{i'=1}^{K} p_{i'j}(\theta_j \mid a_{i'}, b_{i'}, c_{i'}), \tag{4}$$

where j denotes an examinee, i indexes the item for Form X, and i' indexes the item for Form Y. That is, the true score on Form X is considered to be equivalent to that of Form Y given the same ability. This relationship between two true scores is used for 3PL true score equating (Kolen & Brennan 1995):

$$\text{irt}(\tau_X) = \tau_Y(\tau_X^{-1}), \tag{5}$$

where irt represents the IRT true score equating function and τ_X^{-1} is defined as the θ corresponding to true score τ_X.

Beta 4 True Score Equating (B4T)

Suppose that two tests, Form X and Form Y, measure the same psychological trait such that their true scores have a functional relation for any group of examinees. Then, there exists a functional relationship $T_Y = \psi(T_X)$ that is strictly increasing (Lord, 1965). τ_Y, true score for Form Y, can be calculated from

$$\int_{l_X}^{\tau_X} f(\tau)\; d\tau = \int_{l_Y}^{\psi(\tau_X)=\tau_Y} g(\tau)\; d\tau, \tag{6}$$

where $f(\tau)$and $g(\tau)$ are true score distributions of Form X and Form Y, $0 \leq l_X <$ $\tau_X \leq u_X \leq 1$, and $0 \leq l_Y < \tau_Y \leq u_Y \leq 1$. This true score relationship between Form X and From Y is directly applied to observed scores, not to true score.

3PL Observed Score Equating (3PLO)

The first step in IRT observed score equating is to estimate the distributions of observed number-correct scores on Form X and Form Y. For Form X, the conditional distribution of observed scores given ability can be constructed using the Lord and Wingersky (1984) recursion formula. To implement the recursion formula, define X_r(X_r=0, 1, ..., r $\leq$K) as the random variable score on the first r items on the Form X. For r > 1, the recursion formula is as follows:

$$\begin{aligned} f(X = i \mid \theta) &= f(X_{r-1} = i \mid \theta)[1 - p_r(\theta)], & i=0, \qquad (7) \\ &= f(X_{r-1} = i \mid \theta)[1 - p_r(\theta)] + f(X_{r-1} = i-1 \mid \theta)p_r(\theta), & 0< i <r, \\ &= f(X_{r-1} = i-1 \mid \theta)p_r(\theta), & i=r. \end{aligned}$$

These conditional observed score distributions are cumulated over the population to generate the unconditional number-correct distribution of Form X, $f(x)$ (Kolen & Brennan 1995). For Form Y, the same procedure is applied to generate an unconditional number-correct distribution, $g(y)$. The second step is to conduct conventional equipercentile equating on these estimated distributions.

Beta 4 Observed Score Equating (B4O)

Beta 4 observed score equating is conducted using the two steps: first, use beta 4 to fit the observed score distributions of Form X and Form Y. Second, apply conventional equipercentile equating to these fitted distributions. Unlike beta 4 true score equating, there is no equating range restriction in beta 4 observed score equating, because observed scores are used in equating. In the first step, the first four central moments (mean, standard deviation, skewness, kurtosis) of the fitted distributions usually equal those of the observed score distributions. In some cases, only the first three moments, or the first two moments of the fitted distributions equal those of the observed score distributions. So, fitted distributions of Form X and Form Y can have a different number of fitted moments.

Equating Criteria

The following four equating properties were employed as criteria for this study: first-order equity (Morris, 1982), second-order equity (Morris, 1982), conditional-mean-squared-error difference (Thomasson, 1993), and the equipercentile equating property (Angoff, 1971).

Departures from first-order equity and second-order equity of an equating method at each true score can be investigated using the following first-order equity (FOE) bias and second-order equity (SOE) deviation:

$$\text{FOE bias} = E[eq_Y(X) \mid \psi(T_X) = \tau] - E(Y \mid T_Y = \tau\ ) \text{ for all } \tau . \qquad (8)$$

$$\text{SOE deviation} = \sigma^2\,[eq_Y\ (\text{X})|\ \psi(T_X) = \tau\,] - \sigma^2\,(\text{Y}|\,T_Y = \tau\,) \text{ for all } \tau . \qquad (9)$$

In IRT, above two equations can be expressed by replacing τ with θ:

$$\text{FOE bias} = E[eq_Y(X) \mid \theta] - E(Y \mid \theta) \text{ for all } \theta . \qquad (10)$$

$$\text{SOE deviation} = \sigma^2\,[eq_Y\ (\text{X})|\ \theta = \tau\,] - \sigma^2\,(\text{Y}|\,\theta\,) \text{ for all } \theta . \qquad (11)$$

Equations 8 and 10 involve two terms: one is the conditional expected value of the Form Y equivalent, and the other is the conditional expected value of the Form Y score. When the two terms coincide, first-order equity is satisfied and FOE bias is equal to 0. When the two terms are different, first-order equity fails by the difference between the terms. Second-order equity, which quantifies measurement error, is satisfied when the variances of two conditional distributions are the same.

When first-order equity is not satisfied over all true scores or not constant in a specified range of true scores, second-order equity should be considered together with first-order equity (Thomasson, 1993). The conditional-mean-squared-error (CMSE) difference is defined as following:

$$\text{CMSE difference} = \sigma^2\,[eq_Y\ (\mathrm{X})|\ \psi(T_X) = \tau\,] - \sigma^2\,(\mathrm{Y}|T_Y = \tau\,) + \qquad (12)$$

$$\{\,E[eq_Y(X)\,|\,\psi(T_X) = \tau] - E(Y\,|\,T_Y = \tau\ \ )\}^2 \text{ for all } \tau\,.$$

$$= \text{departure from second-order equity} + (\text{bias})^2$$

The psychometric methodology used to calculate these three criteria were given by Kolen, Hanson, and Brennan (1992) and Kolen, Zeng, and Hanson (1996).

Two overall indexes, the unweighted root mean square differences (URMS) and the weighted root mean square differences (WRMS), were computed for these three criteria. The URMS of first-order equity is

$$[\ \int_{-4}^{4} \mathrm{U}(\theta)\ \{\text{FOE bias }\}^2\ d\theta\ \ ]^{1/2} \text{ in IRT, or} \qquad (13)$$

$$[\ \int_{l_\tau}^{u_\tau} \mathrm{U}(\tau\,)\ \{\text{FOE bias }\}^2\ d\tau\ \ ]^{1/2} \text{ in beta 4,} \qquad (14)$$

where U(θ) and U(τ) represent uniform distributions.

The WRMS of first-order equity is

$$[\ \int_{-4}^{4} \psi(\theta)\ \{\text{FOE bias}\}^2\ d\theta\ \ ]^{1/2} \text{ in IRT, or} \qquad (15)$$

$$[\ \int_{l_Y}^{u_Y} \psi(\tau)\ \{\text{FOE bias}\}^2\ \ d\tau\ \ ]^{1/2} \text{ in beta 4,} \qquad (16)$$

where $\psi(\theta)$ is the ability distribution and $\psi(\tau)$ is the true score distribution of Form Y. The similar definition is applied to URMS and WRMS for SOE and CMSE difference.

To evaluate the relative accuracy of the equipercentile equating property, the following sum of absolute differences (SAD) was computed:

$$\text{SAD} = \int_{y} \left| G^{*}[eq_{Y}(x) = y] - G(y) \right| \, dy \,, \tag{17}$$

where $G^{*}[eq_{Y}(x) = y]$ represents the cumulative distribution of a score y in the population of Form Y equivalents.

A Procedure for Comparing Equity Properties

A direct comparison of equity properties for beta 4 equating and 3PL equating is not possible because of their different underlying psychometric models. For comparing FOE biases of beta 4 equating and 3PL equating, the conversion table for beta 4 equating can be used under the assumption that the underlying psychometric model is 3PL: this procedure can be performed simply by replacing $eq_{Y}(X)$ in Equation 10 (equating function of 3PL equating) with $eq_{Y}(X)$ of beta 4. The same procedure is also possible for the IRT equating procedure, by replacing $eq_{Y}(X)$ in Equation 8 (equation function of beta 4 equating) with $eq_{Y}(X)$ of 3PL. That is, the first-order equity properties for beta 4 equating and 3PL equating can be compared on the true score scale of beta 4, or the true score scale of 3PL. The procedure is the same for the comparison of second-order equity and CMSE difference.

Data and Method

Data

Four subtests of the Iowa Test Basic Skills (ITBS; Hoover, Hieronymus, Frisbie, & Dunbar, 1994) battery were used in this study: Vocabulary, Science, Math Problem Solving

Table 1

Descriptive Statistics for Data Sources Used in This Study

Test	V4		S4		M8		MD8	
Form	X	Y	X	Y	X	Y	X	Y
No. of items	34	34	35	35	36	36	33	33
No. of stimulus material					8 (8 4 4 4 4 4 4 4)	8 (8 4 4 4 4 4 4 4)	5 (7 7 6 6 7)	5 (7 7 6 6 7)
Mean	18.5803	18.6310	16.1176	17.7706	18.2716	16.2256	15.3450	15.8217
SD	7.4159	6.9828	6.5842	6.5352	6.5798	6.2490	6.1326	6.2570
Skewness	0.0733	0.0045	0.3459	0.0286	0.2726	0.4468	0.5254	0.3297
Kurtosis	2.0817	2.1610	2.2707	2.0079	23648	2.4891	2.6158	2.3426
KR-20	0.8833	0.8695	0.8348	0.8342	0.8406	0.8160	0.8178	0.8265
Feldt					0.8249	0.8007	0.8134	0.8233

Note. V4=Grade 4 Vocabulary, S4=Grade 4 Science, M8= Grade 8 Math Problem Solving and Data Interpretation, MD8=Grade 8 Maps and Diagrams. Feldt: reliability based on Feldt's average error variance (1984).

and Data Interpretation (M), and Maps and Diagrams (MD). Vocabulary may be the most unidimensional test in the ITBS test battery. In addition, because items on this test are discrete items and associated with common stimuli, alternate forms are very similar in difficulty. Science consists of rather difficult items. M consists of dichotomous items and several stimulus materials, and MD consists of several stimulus materials. The M and MD tests are more likely to violate the unidimensionality assumption of IRT than are the other tests. The data for this study were taken from the ITBS Form K and Form L national standardization process. A Form L (new form, or Form X) test was equated to a Form K (old form, or Form Y) test under a random groups design. For each test, a random sample of 3000 examinees was obtained.

Summary of Equating Procedures

3PL equating methods were conducted following the procedures described by Kolen & Brennan (1995), and Beta 4 equating methods were done following the procedures described by Lord (1965, 1981). For 3PL equating, item parameters were estimated using BILOG (Mislevy & Bock, 1990) with the "float" option. For Form X of MD8, the "TPRior" option was used instead of the "float" option to solve a convergence problem. $\psi(\theta)$ was estimated in 40 equally spaced quadrature points specified by BILOG. In 3PL true score equating, Kolen's (1981) ad hoc procedure was used for the scores below the limits set by the item guessing parameters ($\tau \leq \sum_{i=1}^{K} c_i$) and for the all-correct score, because this procedure is simpler than Lord's (1980) procedure. For beta 4 equating, four parameters and $\psi(\tau)$ were estimated using the method of moments (Lord, 1965). In beta 4 true score equating, Kolen's ad hoc procedure was also applied to the scores below the lower limit (l) and scores above the upper limit (u).

Results

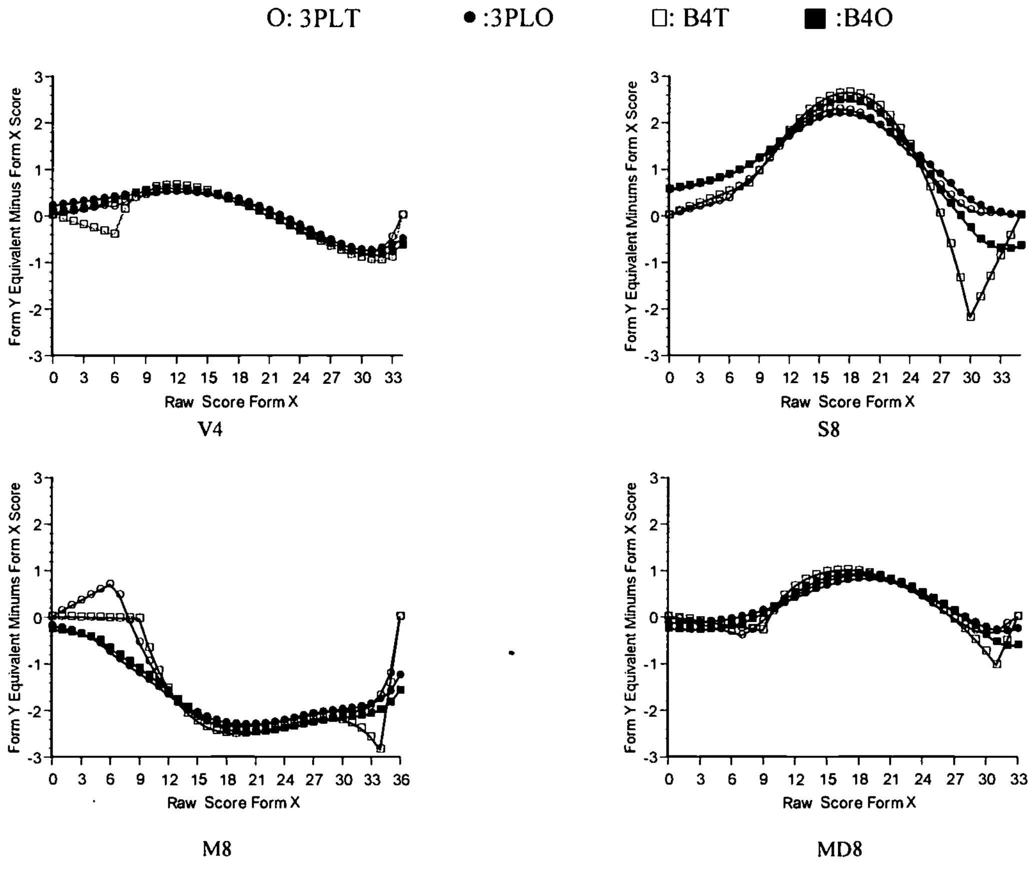
"3PL equating" designates both 3PL true score equating (3PLT) and 3PL observed score equating (3PLO), and "beta 4 equating" designates both beta 4 true score equating (B4T) and beta 4 observed score equating (B4O). "True score equating" refers to both 3PLT and B4T, and "observed score equating" means both 3PLO and B4O. Whenever a comparison between true score equating and observed score equating is mentioned without reference to specific equating methods, this always refers to comparisons of both 3PLT vs. 3PLO and B4T vs. B4O. Whenever a comparison between 3PL equating and beta 4 equating is mentioned without reference to specific equating methods, this always refers to comparisons of both 3PLT vs. B4T and 3PLO vs. B4O.

Equating Results

Figure 1 shows equating results for grade 4 Vocabulary (V4), grade 4 Science (S4), grade 8 Math Problem Solving and Data Interpretation (M8), grade 8 Maps and Diagrams (MD8). To compare equating methods, difference plots are used in this study. The vertical axis is the difference between a Form Y equivalent and a Form X score. To find the Form Y equivalent of a Form X score, add the vertical axis value to the horizontal axis value. If differences between Form X scores and their Form Y equivalents are 0 across all Form X scores, we have identity equating. In general, when two tests are nearly parallel, equating results will be close to the identity equating.

When two test forms are similar in difficulty, 3PL equating and beta 4 equating tended

Figure 1
Equating Results for V4, S4, M8, & MD8



V4

S8

M8

MD8

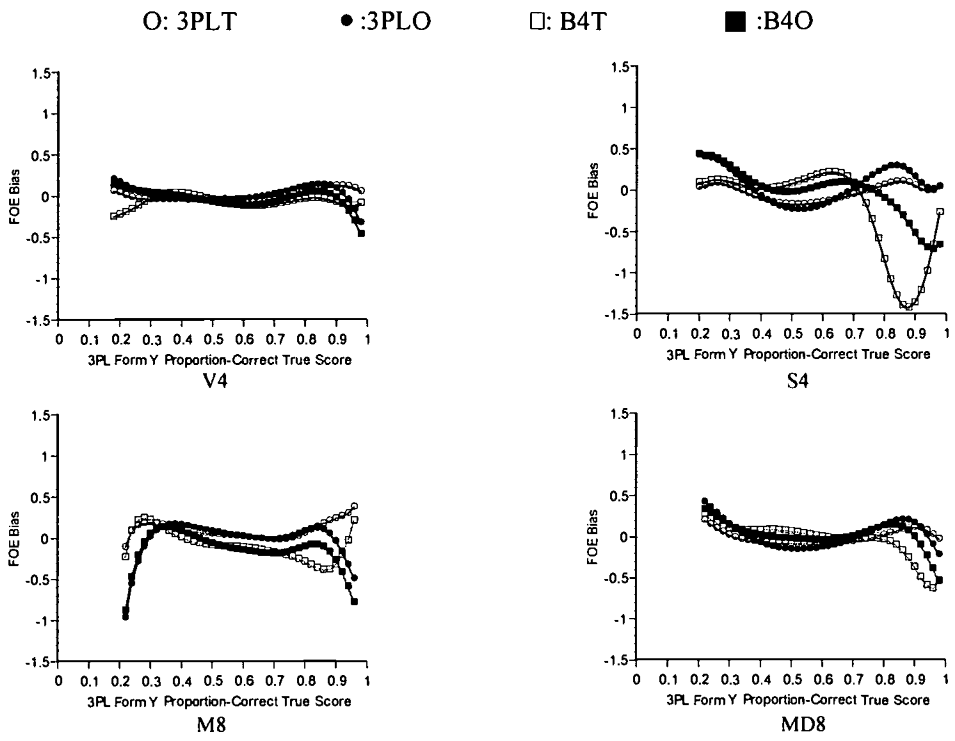
to produce similar equating results. Among the four tests used in this study, the mean form difference in difficulty was less than one raw score point for two tests, V4 and MD8. It is believed that MD8 violates the unidimensional assumption because this test consists of several stimulus materials. As equating relationships for all equating methods for V4 were similar, equating relationships for all equating methods for MD8 were also similar.

The equating relationships between true score equating and observed score equating were often different at Form X low or high scores. For 3PL equating, the largest differences often occurred around the sum of the c parameter estimates and at very high scores, which are near the regions of the score scale where true scores are not achieved. For beta 4 equating, the largest differences occurred around the lower limit (l) and upper limit (u), which are also near the regions of the score scale where true scores are not achieved. The equating relationship for observed score equating might be doubtful at Form X low or high scores because a small number of examinees in these ranges might lead to unstable estimated observed score distributions. The equating relationships between 3PLT and B4T were also often different at these scores. The equating relationships between two observed score equating methods were often more similar than those between two true score equating methods.

First-Order Equity (FOE)

Figure 2 shows FOE bias when the 3PL is the underlying psychometric model. Note that for convenience of interpretation, ability θ was transformed to a 3PL proportion-correct true score using the relationship between true score and θ (see Equation 4). This transformation does not change the pattern of equity properties based on the θ scale. For V4, FOE was nearly satisfied in the sense that FOE biases for all equating methods were close to

Figure 2
First-Order Equity based on the 3PL Model



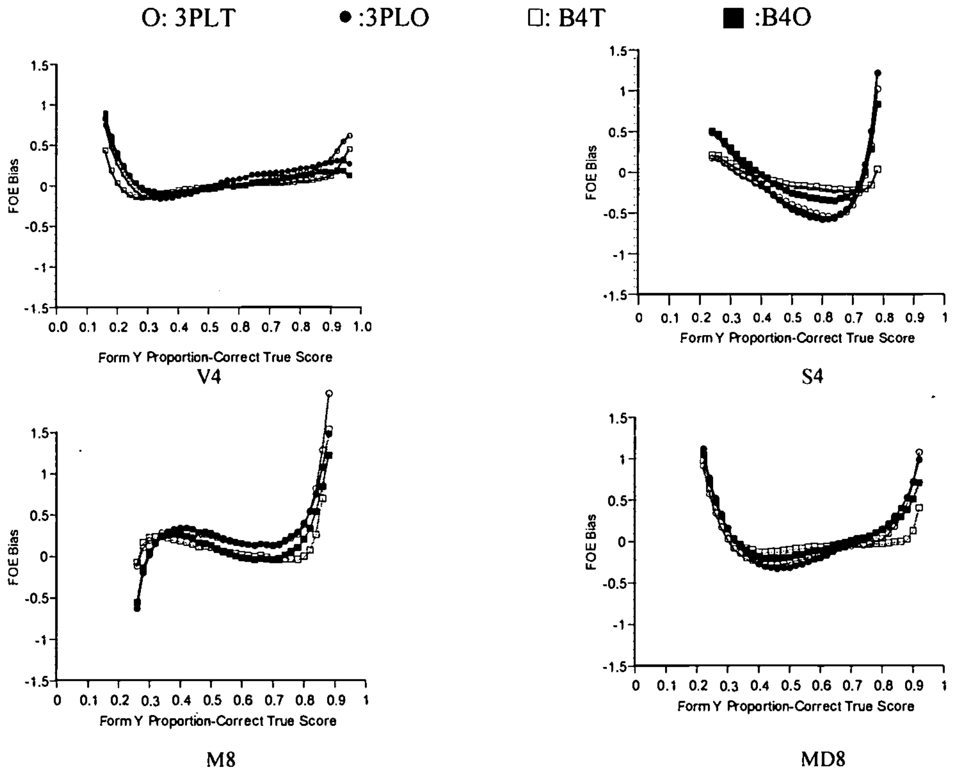
V4

S4

M8

MD8

Figure 3
First-Order Equity based on the beta 4 Model



V4

S4

M8

MD8

0 across all levels of true score. The bias for all equating methods was within ± 0.5. This implies that at any true score, the expected converted score from any one of the four equating methods is within ± 0.5 number-correct score points of a Form Y expected score. For S4, B4T showed relatively large negative bias at true score in the range of 0.8 to 0.95. Figure 3 shows FOE bias when the beta 4 is the underlying psychometric model. Note that the FOE bias was plotted from the lower limit (l) to upper limit (u) of the Form Y true score distribution. For most tests, FOE bias was small in the middle of the true score, but relatively large at low or high true scores, whether the 3PL or the beta 4 is the underlying psychometric model.

The upper part of Table 2 presents the values for the URMS and WRMS of FOE when the 3PL is the underlying psychometric model, while the lower part provides the URMS and WRMS of FOE when the beta 4 is the underlying psychometric model. When the assumptions of an equating method correspond to the underlying psychometric model, first-order equity is more likely to be satisfied than for other equating procedure whose assumptions are different from the underlying psychometric model. This relationship is clear for 3PLT and B4T; 3PLT produces smaller values for the URMS and WRMS of FOE than B4T, when the 3PL model is the underlying psychometric model; B4T provides smaller values for the URMS and WRMS

Table 2
Overall Indexes for First-Order Equity

Common	Equating	URMS				WRMS			
Model	Methods	V	SC	M	MD8	V	SC	M	MD8
3PL	3PLT	**0.06**	**0.08**	**0.23**	**0.12**	**0.05**	**0.11**	**0.11**	**0.08**
	3PLO	0.16	0.27	0.66	0.26	0.07	0.20	0.21	0.15
	B4T	0.15	0.60	0.32	0.21	0.09	0.29	0.15	0.09
	B4O	0.19	0.40	0.66	0.26	0.08	0.17	0.20	0.11
Beta4									
	3PLT	0.21	0.37	0.42	0.29	0.15	0.36	0.23	0.26
	3PLO	0.21	0.42	0.38	0.34	0.16	0.41	0.25	0.32
	B4T	**0.11**	**0.17**	**0.24**	**0.21**	**0.08**	**0.16**	**0.14**	**0.22**
	B4O	0.16	0.25	0.27	0.27	0.11	0.28	0.18	0.26

of FOE than 3PLT, when the beta 4 model is the underlying psychometric model.

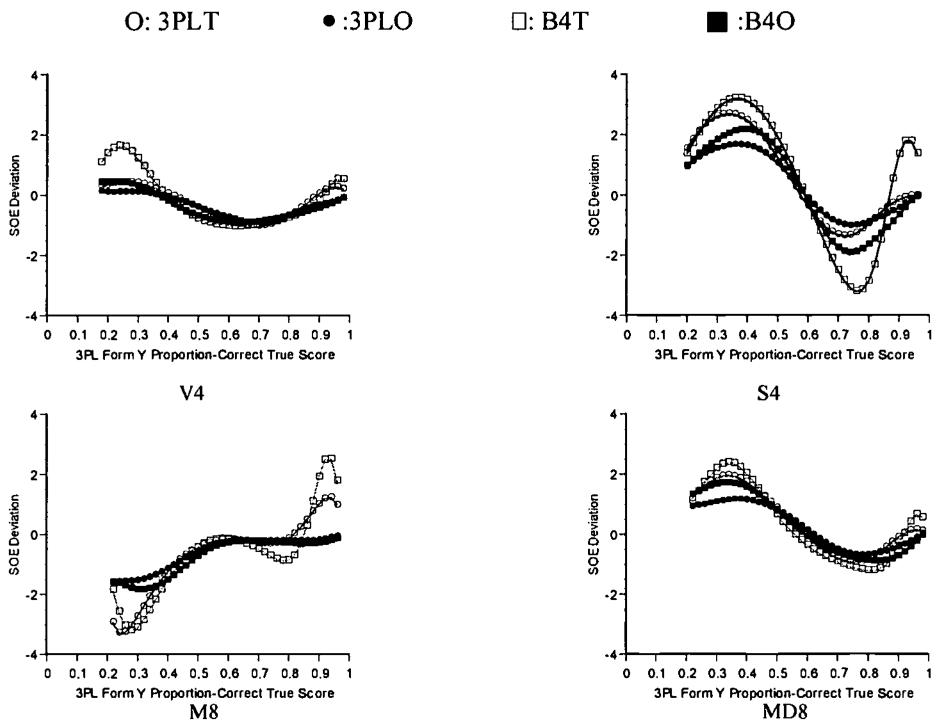
True score equating usually produces smaller first-order equity bias across most levels of true score than observed score equating. When the underlying psychometric model corresponds to the assumptions of equating methods, this relationship clearly appears. Compared to B4O, B4T produces small values of URMS and WRMS for first-order equity when the beta 4 model is the underlying psychometric model. The same conclusions are evident for 3PL equating. 3PLT provides smaller overall index values for first-order equity than 3PLO when the 3PL model is the underlying psychometric model. These results can be predicted, because true score equating is constructed by using true scores, and first-order equity refers to the true score relationship.

Second-Order Equity (SOE) and CMSE difference

Figure 4 shows SOE deviation when the 3PL is the underlying psychometric model, while Figure 5 presents SOE deviation when the beta 4 is the underlying psychometric model. For all equating methods, SOE was also nearly satisfied across most true scores for V4. It was evident that observed score equating produced smaller SOE deviation than true score equating at most true scores, whether the 3PL or the beta 4 is the underlying psychometric model.

Table 3 presents the values for the URMS and WRMS of SOE when the 3PL model and beta 4 model are the underlying psychometric models. Unlike first-order equity, the satisfaction of SOE is not much influenced by whether the underlying psychometric model corresponds to the assumptions of an equating method. Except for a few cases, whether the underlying psychometric model was the 3PL or the beta 4 model, 3PLT produced smaller

Figure 4
Second-Order Equity based on the 3PL Model



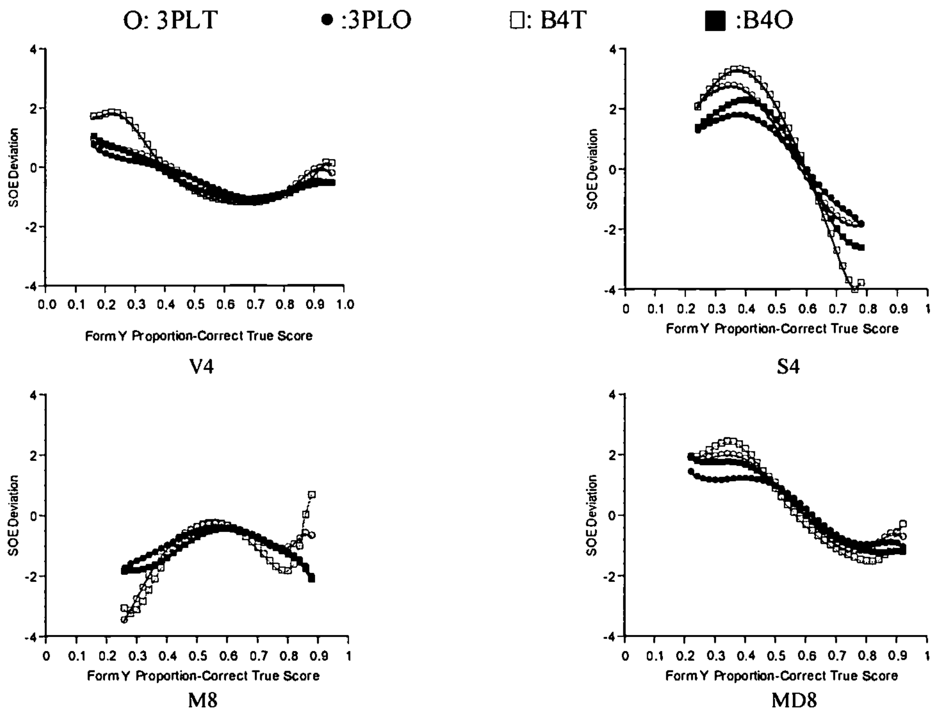
V4

S4

M8

MD8

Figure 5
Second-Order Equity based on the Beta 4 Model



V4

S4

M8

MD8

Table 3
Overall Indexes for Second-Order Equity

Common	Equating	URMS				WRMS			
Model	Methods	V4	S4	M8	MD8	V4	S4	M8	MD8
3PL	3PLT	0.45	1.60	1.98	1.07	0.63	1.77	1.77	1.33
	3PLO	**0.39**	**0.98**	**1.18**	**0.76**	**0.57**	**1.16**	**1.10**	**0.87**
	B4T	0.93	2.01	1.84	1.22	0.90	2.37	1.86	1.53
	B4O	0.50	1.35	1.24	1.08	0.66	1.60	1.31	1.20
Beta4									
	3PLT	0.73	1.89	1.30	1.29	0.76	1.88	1.63	1.44
	3PLO	**0.68**	**1.30**	**1.06**	**0.94**	**0.70**	**1.29**	**1.06**	**0.97**
	B4T	1.06	2.58	1.56	1.50	1.03	2.54	1.82	1.64
	B4O	0.78	1.80	1.24	1.23	0.80	1.76	1.29	1.31

Table 4
Overall Indexes for CMSE difference

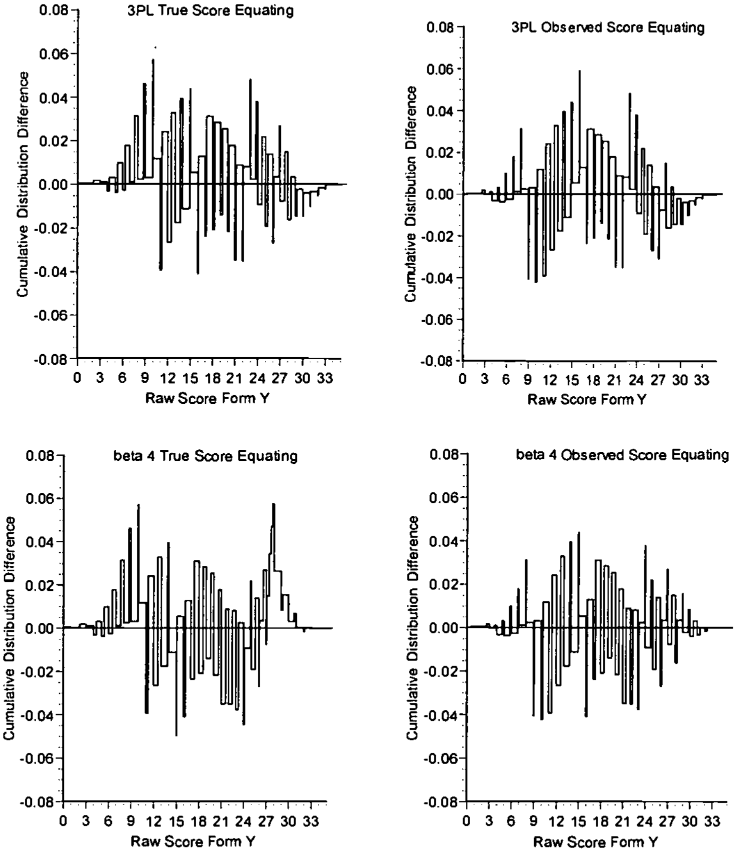
Common	Equating	URMS				WRMS			
Model	Methods	V4	S4	M8	MD8	V4	S4	M8	MD8
3PL	3PLT	0.45	1.53	1.95	1.08	0.63	1.77	1.76	1.33
	3PLO	**0.38**	**1.04**	**0.78**	**0.90**	**0.57**	**1.19**	**1.04**	**0.89**
	B4T	0.95	2.18	1.79	1.24	0.90	2.35	1.84	1.54
	B4O	0.49	1.35	0.93	1.13	0.66	1.63	1.28	1.21
Beta4									
	3PLT	0.74	1.89	1.32	1.34	0.76	1.88	1.60	1.51
	3PLO	**0.68**	**1.33**	**0.91**	**1.01**	**0.69**	**1.31**	**0.99**	**1.09**
	B4T	1.08	2.58	1.58	1.54	1.04	2.54	1.81	1.70
	B4O	0.80	1.80	1.15	1.28	0.80	1.77	1.25	1.40

values of URMS and WRMS for SOE than did B4T, and 3PLO provided smaller values of URMS and WRMS for SOE than did B4O. Because CMSE difference is mainly influenced by SOE, this relationship also applied to CMSE difference. This fact can be found in Table 4. In short, 3PL equating produces smaller values of URMS and WRMS for SOE deviation and CMSE difference than did beta 4 equating. Observed score equating produced smaller SOE deviation and CMSE difference than true score equating. That is, with respect to second-order equity and CMSE difference, 3PLO provides smaller values of URMS and WRMS than 3PLT, B4O provides smaller values than B4T, and among the four equating methods, in general, 3PLO produces the smallest values. An explanation for this result remains to be investigated. This pattern for second-order equity usually appears even when the assumptions

Table 5
Summary Statistics for Equipercentile Equating Property

Equating Methods	Test V4	S4	M8	MD8
3PLT	0.415	0.397	0.407	0.339
3PLO	0.414	0.380	**0.350**	0.339
B4T	0.411	0.466	0.456	0.319
B4O	**0.402**	**0.367**	0.369	**0.313**

Figure 6
Equipercentile Equating Property for Grade 8 Science



of the underlying psychometric model are not the same as the assumptions of equating methods.

Equipercentile Equating Property

Table 5 presents summary statistics for equipercentile equating property. Observed score equating produced smaller EQAD than true score equating, and beta 4 observed score equating produced the smallest EQAD except for M8. In general, the first four moments of an estimated observed score distribution for beta 4 are the same as those of observed score distribution, because beta 4 usually fits the first four moments of the observed score distribution very well. Therefore, it follows that B4O usually produces smaller EQAD than 3PLO. Even when only three moments can be fit using beta 4, B4O produces smaller values for the EQAD than the other three equating methods. As can be seen in Figure 6 for S4, the main differences in the equipercentile equating property between equating methods usually occurred around Form X low or high scores, where equating relationships between equating methods were usually quite different.

Summary and Discussion

Table 6 presents the summary results for 3PLT, 3PLO, B4T, and B4O using four equating criteria, FOE, SOE, CMSE difference, and equipercentile equating property. The following five main conclusions can be drawn from this table. First, when the assumptions of the equating method corresponded to the underlying psychometric model, first-order equity was more likely to be satisfied than for other equating methods whose assumptions were different from the underlying psychometric model. Second, when the underlying psychometric model corresponded to the assumptions of an equating method, true score equating produced smaller FOE bias than did observed score equating. That is, 3PLT produced smaller bias than 3PLO, and B4T produced smaller bias than B4O. Third, observed score equating produced smaller SOE deviation than did true score equating. The same results were obtained for CMSE difference. Fourth, 3PL equating produced smaller SOE deviation than did beta 4 equating. Among the four equating methods, 3PLO produced the smallest SOE deviation. Fifth, observed score equating provided smaller EQAD than did true score equating, and B4O tended to produce the smallest EQAD value.

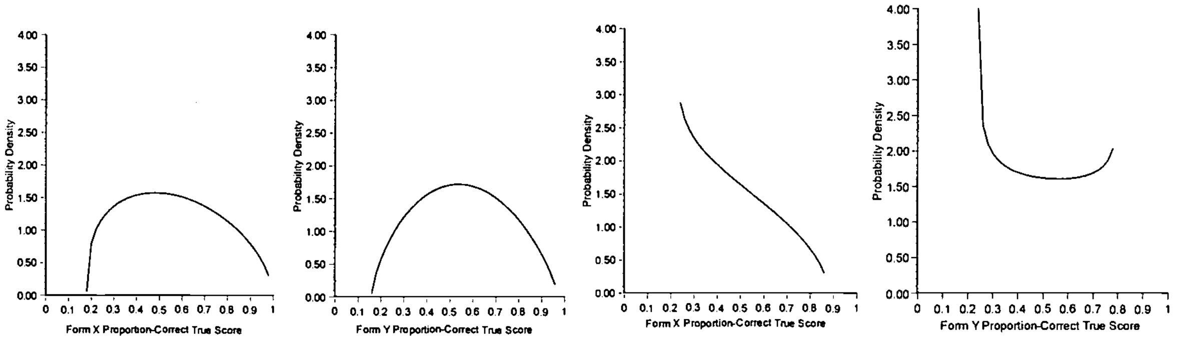
Table 6.
Summary results

Comparison	FOE	SOE and CMSE difference	Equipercentile Equating property
3PLT vs. B4T	**3PLT** based on the 3PL model **B4T** based on the beta 4 model	**3PLT**	Not clear
3PLO vs. B4O	Not clear	**3PLO**	**B4O**
3PLT vs.3PLO	**3PLT**	**3PLO**	**3PLO**
B4T vs. B4O	**B4T** based on the beta 4 model Not clear based on the 3PL model	**B4O**	**B4O**

* **Boldface** represents an equating method that produces smaller values of URMS and WRMS for equity properties or smaller values of EQAD for the equipercentile equating property.

Beta 4 true score equating often produced quite different equating results from those of the other three equating methods. These different equating results were especially apparent at low or high scores. Even when the form difficulty was very small (V4), compared to the other equating methods, B4T produced a slightly different equating relationship at Form X low scores and different equity properties at low true scores. One reasonable explanation might be the possibility of an unstable parameter estimation procedure. The "method of moments" procedure is used to estimate the four parameters (α, β, l, u) of the true score distribution (Lord, 1965; Hanson, 1991). When a test form is difficult, the probability density at the lower limit (l) was usually abnormally large. Figure 7 shows the estimated true score distributions of the beta 4 model for V4 and S4. While the probability densities for Form X and Form Y for V4 were large in the middle of true scores and small at low and high true scores, for S4 which consists of difficult items, the probability density for the lower limit was very large for both Form X and Form Y. That is, there is a possibility that the estimated lower

Figure 7
Estimated True Score Distributions of the Beta 4 Model for V4 and S4



Form X for V4 Form Y for V4 Form X for S4 Form Y for S4

limit of beta 4 could be unstable. It follows that the equating relationship for beta 4 true score equating can be unstable, too. Therefore, it may be useful to investigate the true score distributions for Form X and Form Y, whenever B4T is applied.

In practice, researchers need to carefully investigate first-order equity at low and high true scores because large first-order equity biases often occur in these ranges. Certainly, a bias around ± 1 should be considered to be a problem because a one-number-correct difference often leads to a different scale score, which is reported to test takers or users. For example, one raw score difference leads to a different developmental standard score of ITBS because the raw score and developmental standard score have a one-to-one relationship. When probability densities are very sparse in those true score ranges, however, these biases may not be a serious problem in practice.

If a testing program routinely applies the 3PL model to construct test forms, the results for equity properties might support the use of 3PL equating methods. If first-order equity is considered an important equating criterion, 3PLT can be recommended, and if second-order equity or CMSE difference is important, 3PLO can be suggested. Even when a test form is built based on classical test theory, 3PLO can be considered when second-order equity or CMSE difference is considered an important criterion.

When the equipercentile equating property is considered an important equating criterion, the results from this study suggest that, B4O should be considered. B4O usually closely meets the equipercentile equating property than do 3PLT and 3PLO because of the moment preservation property of beta 4.

Reference

Angoff, W.H. (1971). Scales, norms and equivalent scores. In R. L. Thorndike(Ed.), Educational measurement (2nd ed., pp. 508-600). Washington, DC: American Council on Education.

Cook, L.L., & Eignor, D.R., & Schmitt, A.P. (1988). The effects on IRT and conventional achievement test equating results of using equating samples matched on ability (RR-88-52). Princeton, NJ: Educational Testing Service.

Cook, L.L., & Eignor, D.R. (1991). An NCME instructional module on IRT equating methods. Educational Measurement: Issues and Practice, 10, 37-45.

Eignor, D.R., Stocking, M. L., & Cook, L.L. (1986). Simulation results of effects on linear and curvilinear observed- and true-score equating procedures of matching on a fallible criterion. Applied Measurement in Education, 3, 37-52.

Eignor, D.R., Stocking, M. L., & Cook, L.L. (1995). The effects on observed- and true-score equating procedures of matching on a fallible criterion: a simulation with test variation (RR-90-25). Princeton, NJ: Educational Testing Service.

Felt, L.S. (1984). Some relationships between the binomial error model and classical test theory. Educational and Psychological Measurement, 44, 883-891.

Han, T., Kolen, M.J., & Pohlmann, J. (1997). A comparison among IRT true-and observed score equating and traditional equipercentile equating. Applied Measurement in Education, 10, 105-121.

Hanson, B. A. (1991a). A note on Levine's formula for equating unequally reliable tests using data from the common item nonequivalent groups design. Journal of Educational Statistics, 16, 93-100.

Harris, D. J., & Kolen, M.J. (1986). Effect of examinee group on equating relationships. Applied Psychological Measurement, 10, 35-43.

Hoover, H.D., Hieronymus, A., Frisbie, D., and Dunbar, S. (1994). Iowa Test Basic Skills: Interpretive Guide for School Administrations. Chicago, IL: The Riverside Publishing Company

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

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

Kolen, M.J., Hanson, B.A., & Brennan, R.L. (1992). Conditional standard errors of measurement for scale scores. Journal of Educational Measurement, 29, 285-307.

Kolen, M.J., & Harris, D.J. (1990). Comparison of item preequating and random groups equating using IRT and equipercentile methods. Journal of Educational Measurement, 27, 27-39.

Kolen, M.J., Zeng, L., & Hanson, B.A. (1996). Conditional standard errors of measurement for scale scores using IRT. Journal of Educational Measurement, 33, 129-140.

Lawrence, I.M., & Dorans, N.J. (1990). Checking the statistical equivalence of nearly identical test editions. Applied Measurement in Education, 3, 245-254.

Livingston, S.A., Dorans, N.J., & Wright, N.K. (1990). What combination of sampling and equating methods works best? Applied Measurement in Education, 3, 73-95.

Lord, F.M. (1965). A strong true score theory with applications. Psychometrika, 30, 239-270.

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

Lord, F.M., & Novick, M.R. (1968). Statistical theories of mental test score. Menlo Park, CA:Addison-Wesley.

Lord, F.M., & Wingersky, M.S. (1984). Comparisons of IRT true score and equipercentile observed-score "equating." Applied Psychological Measurement, 8, 453-461.

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

Morris, G. N. (1982). On the foundations of test equating. In P.W. Holland & D.B. Rubin (Eds.), Test equating (pp. 169-191). New York: Academic Press.

Skaggs, G. (1990). To match or not to match samples on ability for equating: A discussion of five articles. Applied Measurement in Education, 3, 105-113.

Skaggs, G., & Lissitz, R.W. (1986). An exploration of robustness of four test equating models. Applied Psychological Measurement, 10, 303-317.

Thomasson, G. L. (1993). The asymptotic equating methodology and other test equation evaluation procedure. Unpublished doctoral dissertation, University of Illinois at Urbana-Champaign.

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

ERIC®

TM033765

REPRODUCTION RELEASE

(Specific Document)

I. DOCUMENT IDENTIFICATION:

Title: A Comparison of IRT Equating and Beta 4 Equating

Author(s): Dong-In Kim, Robert Brennan, Michael Kolen

Corporate Source:

Publication Date:

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic media, and sold through the ERIC Document Reproduction Service (EDRS). Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce and disseminate the identified document, please CHECK ONE of the following three options and sign at the bottom of the page.

The sample sticker shown below will be affixed to all Level 1 documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

1

Level 1

☑

Check here for Level 1 release, permitting reproduction and dissemination in microfiche or other ERIC archival media (e.g., electronic) *and* paper copy.

The sample sticker shown below will be affixed to all Level 2A documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE, AND IN ELECTRONIC MEDIA FOR ERIC COLLECTION SUBSCRIBERS ONLY, HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

2A

Level 2A

□

Check here for Level 2A release, permitting reproduction and dissemination in microfiche and in electronic media for ERIC archival collection subscribers only

The sample sticker shown below will be affixed to all Level 2B documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

2B

Level 2B

□

Check here for Level 2B release, permitting reproduction and dissemination in microfiche only

Documents will be processed as indicated provided reproduction quality permits.
If permission to reproduce is granted, but no box is checked, documents will be processed at Level 1.

I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce and disseminate this document as indicated above. Reproduction from the ERIC microfiche or electronic media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries.

Sign here,→ please

Signature:	Printed Name/Position/Title: Kim, Dong-In	
Organization/Address: CTB/McGrawHill, 20 Ryan Ranch Road Monterey, CA 93940	Telephone: 831-393-7421	FAX: 831-393-7016
	E-Mail Address: dKim@ctb.com	Date: 4/1/02

(over)

III. DOCUMENT AVAILABILITY INFORMATION (FROM NON-ERIC SOURCE):

If permission to reproduce is not granted to ERIC, *or*, if you wish ERIC to cite the availability of the document from another source, please provide the following information regarding the availability of the document. (ERIC will not announce a document unless it is publicly available, and a dependable source can be specified. Contributors should also be aware that ERIC selection criteria are significantly more stringent for documents that cannot be made available through EDRS.)

Publisher/Distributor:

Address:

Price:

IV. REFERRAL OF ERIC TO COPYRIGHT/REPRODUCTION RIGHTS HOLDER:

If the right to grant this reproduction release is held by someone other than the addressee, please provide the appropriate name and address:

Name:

Address:

V. WHERE TO SEND THIS FORM:

Send this form to the following ERIC Clearinghouse:

University of Maryland
ERIC Clearinghouse on Assessment and Evaluation
1129 Shriver Laboratory
College Park, MD 20742
Attn: Acquisitions

However, if solicited by the ERIC Facility, or if making an unsolicited contribution to ERIC, return this form (and the document being contributed) to:

ERIC Processing and Reference Facility
4483-A Forbes Boulevard
Lanham, Maryland 20706

Telephone: 301-552-4200
Toll Free: 800-799-3742
FAX: 301-552-4700
e-mail: ericfac@inet.ed.gov
WWW: http://ericfac.piccard.csc.com