CRESST REPORT 751

Taehoon Kang
Troy T. Chen

AN ALTERNATIVE IRT OBSERVED SCORE EQUATING METHOD

FEBRUARY, 2009

National Center for Research on Evaluation, Standards, and Student Testing

Graduate School of Education & Information Studies
UCLA | University of California, Los Angeles

An Alternative IRT Observed Score Equating Method

CRESST Report 751

Taehoon Kang
CRESST/University of California, Los Angeles

Troy T. Chen
ACT, Inc.

February, 2009

National Center for Research on Evaluation,
Standards, and Student Testing (CRESST)
Center for the Study of Evaluation (CSE)
Graduate School of Education & Information Studies
University of California, Los Angeles
300 Charles E. Young Drive North
GSE&IS Building, Box 951522
Los Angeles, CA 90095-1522
(310) 206-1532

The work reported herein was supported under the National Research and Development Centers, Award Number R305A050004, as administered by the U.S. Department of Education's Institute of Education Sciences (IES).

AN ALTERNATIVE IRT OBSERVED SCORE EQUATING METHOD

Taehoon Kang
CRESST/University of California, Los Angeles

Troy T. Chen
ACT, Inc.

Abstract

In this report, an alternative item response theory (IRT) observed score equating method was newly developed. The proposed equating method was illustrated with two real data sets and the equating results were compared to those of traditional IRT true score and IRT observed score equating methods. Using three loss indices, the new method appeared to produce equating equivalents more similar to those of the IRT observed score equating than those of the IRT true score equating. In addition to the conversion relationships between new form scores and their equating equivalents on the old form scale, the bootstrap standard errors of equating were provided and compared for the three IRT equating methods. These methods performed similarly.

Introduction

The number-correct scores from different test forms often need to be equated for the purpose of evaluating examinees' proficiency across different forms or years. Traditionally, under item response theory (IRT), there are two equating methods to adjust the raw test scores of the new form X onto the old form Y metric (Lord, 1982): IRT true score equating (IRT-TSE) and IRT observed score equating (IRT-OSE). The former discovers the equivalent score on Y metric, $\varphi(x)$, for an observed score x on form X using the test characteristic curves for both forms which respectively define the relationship between person location parameters (i.e., θ) and the corresponding true test scores. The latter depends upon the traditional equipercentile equating method after constructing the expected raw score distributions of two test forms which are typically obtained with the use of the recursive algorithm (Lord & Wingersky, 1984; Thissen, Pommerich, Billeaud, & Williams, 1995).

IRT-OSE has explicit advantages over IRT-TSE because IRT-OSE deals with observed scores of actual interest in addition to the fact that it could be controversial to treat estimated true scores as substitutes for observed scores under IRT-TSE. Also, whereas the IRT-TSE method cannot produce equating equivalents for a perfect score or an observed score of x less than the sum of the guessing parameters under the 3-parameter logistic model, the IRT-OSE method can (Han, Kolen, & Pohlmann, 1997; Harris & Crouse, 1993; Kolen & Brennan,

2004; Lord, 1977). In practical situations, however, IRT-TSE has been widely used as an alternative to IRT-OSE because this method is easier to conduct and does not require the use of any distribution of ability or expected raw scores.

The main goal of this report is to propose an alternative IRT observed score equating method (AIRT-OSE) that does not require relying on the use of classical equipercentile equating method. The newly proposed AIRT-OSE method employs estimated θ values associated with each number-correct score for equating the observed scores on the two test forms. To this end, Thissen and Orlando's (2001) ability estimation method known as *expected a posteriori* under summed scoring (EAP_{SS}) is used. In this report, the results of the AIRT-OSE method are compared to those of the traditional IRT-OSE and IRT-TSE methods. The next section begins with an introduction to EAP_{SS} and details the AIRT-OSE procedure.

The AIRT-OSE method

Expected a posteriori under summed scoring (EAP_{SS})

For item i on an I-item test, let $u_i = 1$ if an examinee responded correctly and $u_i = 0$ otherwise. Let $\hat{\theta}_{x=\sum u_i}$ denote the EAP_{SS} values for students being administered form X. According to Thissen and Orlando (2001), the EAP_{SS} for a student who earned a raw score x on form X is given by

$$\hat{\theta}_{x=\sum u_i} = \frac{\int \theta L_x(\theta) d\theta}{\int L_x(\theta) d\theta} \approx \frac{\sum_q \theta_q L_x(\theta_q)}{\sum_q L_x(\theta_q)}, \qquad (1)$$

where $L_x(\theta_q)$ is the likelihood for each score x at a given quadrature point θ_q. The likelihood $L_x(\theta_q)$ in Equation 1 can be computed as follows:

$$L_x(\theta_q) = \sum_{\mathbf{u} \in x} \prod_i^I P_i(\theta_q)^{u_i} [1 - P_i(\theta_q)]^{1-u_i} \phi(\theta_q), \qquad (2)$$

where $\phi(\theta_q)$ represents a discrete distribution emulating the population density, u denotes the item response pattern of an examinee such that $x = \sum u_i$, and $P_i(\theta_q)$ is the probability of the correct response to item i at a given θ_q. Under the IRT model employed for analyzing test data, $P_i(\theta_q)$ is computed. For example, in the case of the 3-parameter logistic model (3PLM), it is

$$P_i(\theta_q) = c_i + (1 - c_i)\frac{1}{1 + \exp[-1.7a_i(\theta_q - b_i)]} \qquad (3)$$

where a_i, b_i and c_i denote the discrimination, difficulty and guessing parameter estimates, respectively. Under the EAP_{SS} approach, consequently, the ability estimate will be the same for all students having the same raw score on form X regardless of their response patterns. The EAP_{SS} for a student who got the number-correct score y on form Y (i.e., $\hat{\theta}_{y=\sum u_i}$) can be computed using the same procedure.

The AIRT-OSE procedure

Among various linking designs are the common-item-test design and random-equivalent-group design. For the former, the item parameters can be placed on the same metric via several approaches using a set of common items (e.g., mean-sigma method, test characteristic curve method, concurrent calibrations, etc.). Under a random-equivalent-group design, separate calibrations produce item parameter estimates that are considered to be on the same scale. Once the item parameter estimates of two forms are placed onto a common scale, $\hat{\theta}_{x=\sum u_i}$ and $\hat{\theta}_{y=\sum u_i}$ values calibrated using them are considered to be on the same metric.

When the EAP_{SS} estimates of two forms X and Y are available, the procedure for implementing AIRT-OSE can be summarized as follows:

1. Specify an observed score x on form X.
2. Find the EAP_{SS} ($\hat{\theta}_{x=\sum u_i}$) that corresponds to the observed score x. Let the magnitude of this EAP_{SS} be represented by $\hat{\theta}_*$.
3. Find the equating equivalent, $\varphi(x)$, on the form Y scale that corresponds to $\hat{\theta}_*$.

Typically, most $\varphi(x)$ values resulting from Step 3 will not be whole numbers. This is because the nonlinear relationship between the EAP_{SS} values and the raw scores is one-to-one unique for each form. Thus, to estimate $\varphi(x)$ in Step 3, a few possible interpolation methods are suggested in this report and will be explained following an example detailing the 3-step process presented above.

Assume there are two 40-dichotomous-item test forms X and Y, which share a set of common items. Upon successful calibration of each form, their item parameter estimates are placed on a common scale through, for example, the Stocking and Lord (1983) method. The conversions between the EAP_{SS} and the observed raw scores for each form are then

computed using these item parameter estimates, which are on the same metric. Figure 1 exhibits the plots of the conversions for this illustrative example.

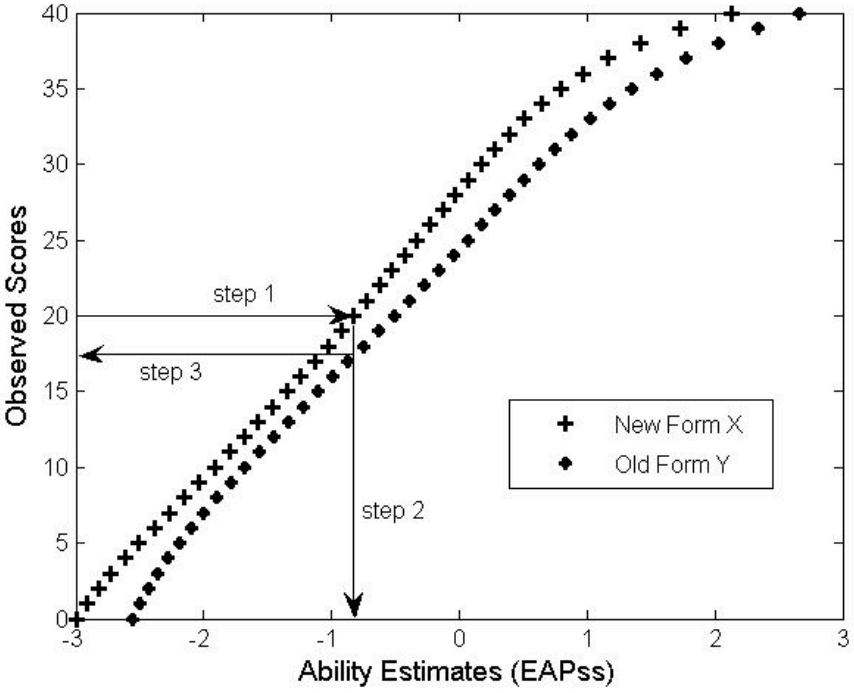

Figure 1. An illustrative equating based on the 3-step process of AIRT-OSE.

In Figure 1, for an examinee having an observed score $x = 20$ on the new form X, the corresponding EAP_{SS} is found to be $\hat{\theta}_* = -.82$ and the corresponding equivalent on the form Y scale, $\varphi(x = 20)$, is found to be between $y = 17$ and $y = 18$. To decide a point-estimate $\varphi(x = 20)$, subsequently, three methods are considered: The first method is a polynomial curve fitting (PCF) approach, the second method is the linear spline interpolation (LSI) approach, and the third method is a cubic spline interpolation (CSI) approach. Following is a discussion of the three approaches.

Under the PCF method, the following n^{th} degree polynomial is employed to fit the score points of the old form Y:

$$y = \beta_n \hat{\theta}^n + \beta_{n-1} \hat{\theta}^{n-1} + \cdots + \beta_1 \hat{\theta} + \beta_0 + \varepsilon \qquad (4)$$

where y and $\hat{\theta}$ are the observed raw score for form Y and the corresponding EAP_{SS}, respectively. The β s represent the fitting coefficients and β_n is expected to be larger than zero. The degree of the polynomial, n, is thought of as an odd integer (e.g., 3, 5, 7, 9, etc.) in

this report. In Equation 4, the β s can be estimated, for example, using the built-in routine, POLYFIT, of the computer software MATLAB (The MathWorks, 2003). For the previous example, the estimate of $\varphi(x = 20)$ is given as follows:

$$\hat{\varphi}(x = 20) = \hat{\beta}_n \hat{\theta}_*^{\,n} + \hat{\beta}_{n-1} \hat{\theta}_*^{\,n-1} + \cdots + \hat{\beta}_1 \hat{\theta}_* + \hat{\beta}_0 .$$

The second approach for estimating $\varphi(x)$ is the linear spline interpolation (LSI). In mathematics, a spline is a function constructed of piecewise polynomial functions. The piecewise functions are connected at the endpoints of contiguous intervals with a certain degree of smoothness for the resulting function. An extensive explanation of spline is provided by de Boor (2001). For LSI, the piecewise function for each interval is linear, which is the simplest spline. Under LSI, $\varphi(x = 20)$ in the interval of (17, 18) for the previous example can be estimated by the following

$$\hat{\varphi}(x = 20) = \frac{\hat{\theta}_* - \hat{\theta}_{y=17}}{\hat{\theta}_{y=18} - \hat{\theta}_{y=17}} \times (18 - 17) + 17 .$$

The third approach for estimating $\varphi(x)$ considered in this report is cubic spline interpolation (CSI). For this report, another MATLAB routine, CSAPS which is based on the cubic spline method introduced in Schoenberg (1964) and Reinsch (1967), is used for implementing CSI. In contrast to LSI which uses linear piece-wise functions, the cubic spline applies third-order polynomial, *S*, within an interval. As de Boor (2001) explained, the function of a cubic spline curve can be obtained by minimizing

$$p \sum_{i=1}^{I} \{y_i - S(\hat{\theta}_i)\}^2 + (1 - p) \int_{\hat{\theta}_{i=1}}^{\hat{\theta}_{i=I}} \{S''(t)\}^2 dt , \qquad (5)$$

where *i* indicates each data point (e.g., $I = 41$ when a test has 40 dichotomous items). The first and second terms in Equation 5 are called the error measure and the roughness measure, respectively, while the degree of smoothness is controlled by the smoothing parameter $p \in [0,1]$. The smaller the *p* value, the smoother the spline. For $p = 0$, the fitted curve will be linear as the ordinary least squares (OLS) line. The choice of $p = 1$ produces an unsmooth curve that passes through all data points. In this report, four levels of smoothness with $p = 1$, 0.75, 0.50, and 0.25 are considered.

For the IRT observed score equating methods (i.e., both IRT-OSE and AIRT-OSE) under examination in this study, an operational rule was applied in deciding $\hat{\varphi}(x)$ values: If $\hat{\varphi}(x)$ is less than zero, it is set to be zero. And, when $\hat{\varphi}(x)$ is higher than the perfect score

of the old form Y, it is assigned the perfect score of form Y. This rule was adopted for the practical reason that IRT-OSE or AIRT-OSE could produce $\hat{\varphi}(x)$ lower than zero or higher than the perfect score. Under AIRT-OSE, for example, in Figure 1, $\hat{\varphi}(x=2)$ was set to zero rather than a negative value. In other words, when an EAP_{SS} estimate $\hat{\theta}_x$ is lower than the minimum of $\hat{\theta}_y$ (i.e., $\hat{\theta}_{y=0}$), the corresponding $\hat{\varphi}(x)$ was set to zero, the lowest raw score. And, when an EAP_{SS} estimate $\hat{\theta}_x$ is higher than the maximum of $\hat{\theta}_y$ (i.e., $\hat{\theta}_{y=perfect}$), the corresponding $\hat{\varphi}(x)$ was set to be the perfect score of form Y.

Evaluation and comparison of the results of the three IRT equating methods

In the following section, the application of the AIRT-OSE method is illustrated with two real test data sets. The results of the method are compared to those of IRT-OSE and IRT-TSE. The comparisons are presented numerically and graphically in terms of the following:

Three loss indices including mean signed difference (MSD), mean absolute difference (MAD), and root mean squared difference (RMSD) in $\hat{\varphi}(x)$ values for two different equating methods (Han, Kolen, & Pohlmann, 1997). Each index is weighted by the frequency of form X scores. The three loss indices are computed as follows:

$$\text{MSD} = \frac{\sum_{x=0}^{X} [\hat{\varphi}(x)_A - \hat{\varphi}(x)_B] f_x}{N}, \qquad (6)$$

$$\text{MAD} = \frac{\sum_{x=0}^{X} | \hat{\varphi}(x)_A - \hat{\varphi}(x)_B | f_x}{N}, \qquad (7)$$

$$\text{RMSD} = \sqrt{\frac{\sum_{x=0}^{X} [\hat{\varphi}(x)_A - \hat{\varphi}(x)_B]^2 f_x}{N}}, \qquad (8)$$

where A and B denote two different IRT equating methods under investigation, f_x is the observed frequency, and N is the sample size of the group that was administered form X.

Patterns of the conversion from form X to form Y scores. Here, $\hat{\varphi}(x) - x$ values are plotted against the raw score x of form X.

The standard errors of equating estimated using the bootstrap method (Efron, 1982; Efron & Tibshirani, 1993; Kolen & Brennan, 2004). From both forms X and Y, 500 random bootstrap samples are drawn, respectively. And, the standard error of equating at a given raw score x is estimated by the standard deviation of the 500 $\hat{\varphi}_r(x)$ values where $r = 1, \ldots, 500$.

Two Illustrations

The three IRT equating methods under investigation were applied to real data for two tests. One is a job skills assessment and the other an academic achievement assessment. Note that results and analyses presented here are only for illustrative purposes and should not be viewed from any other perspective.

Job Skills Assessment

This report used data from two forms of a mathematics job skills assessment consisting of 30 multiple-choice items. The two forms used in the example share 11 common items with each other. One form (Y) was administered to about 3,000 examinees, and the other form (X) was taken by about 1,800 examinees. The data for each form were separately calibrated with the 3-parameter logistic model (3-PLM) using binary logistic models (BILOG; Mislevy & Bock, 1990). The calibrations converged successfully, and the item parameter estimates for form X were placed on the form Y scale using the Stocking and Lord (1983) method.

Table 1

Job Skills Assessment Data: Unrounded Equating Equivalent Estimates, $\hat{\varphi}(x)$ s, from the Three IRT Equating Methods (IRT-TSE, IRT-OSE, and AIRT-OSE).

				AIRT-OSE								
x	f_x	IRT-TSE	IRT-OSE	PCF: 3rd Deg.	PCF: 5th Deg.	PCF: 7th Deg.	PCF: 9th Deg.	LSI	CSI: $p = 1.00$	CSI: $p = 0.75$	CSI: $p = 0.50$	CSI: $p = 0.25$
0	0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
1	1	1.06	0.52	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2	2	2.12	1.09	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
3	8	3.18	1.85	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
4	3	4.24	2.67	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
5	7	5.30	3.52	1.85	1.34	1.35	1.26	1.26	1.27	1.64	1.74	1.81
6	12	5.85	4.25	2.88	3.10	3.15	3.09	3.08	3.09	2.91	2.91	2.94
7	22	6.23	5.08	4.12	4.80	4.83	4.89	4.87	4.88	4.36	4.28	4.26
8	21	6.65	6.03	5.54	6.35	6.34	6.44	6.41	6.43	5.90	5.77	5.73
9	29	7.10	7.12	7.08	7.76	7.71	7.74	7.74	7.75	7.44	7.32	7.27
10	39	7.77	8.36	8.63	9.03	8.98	8.92	8.93	8.93	8.90	8.82	8.80
11	47	8.87	9.63	10.06	10.17	10.13	10.04	10.04	10.04	10.19	10.18	10.18
12	79	10.30	10.86	11.35	11.21	11.20	11.13	11.12	11.12	11.35	11.39	11.41
13	80	11.76	12.07	12.55	12.22	12.23	12.22	12.21	12.20	12.43	12.52	12.56
14	75	13.10	13.25	13.73	13.25	13.28	13.32	13.32	13.32	13.51	13.63	13.68
15	136	14.32	14.42	14.90	14.35	14.39	14.45	14.46	14.46	14.62	14.75	14.81
16	123	15.49	15.57	16.06	15.50	15.54	15.60	15.62	15.62	15.76	15.89	15.93
17	140	16.66	16.73	17.22	16.70	16.72	16.76	16.78	16.78	16.92	17.02	17.05

(table continues)

Table 1 *(continued)*

				AIRT-OSE								
x	*fx*	IRT-TSE	IRT-OSE	PCF: 3rd Deg.	PCF: 5th Deg.	PCF: 7th Deg.	PCF: 9th Deg.	LSI	CSI: $p = 1.00$	CSI: $p = 0.75$	CSI: $p = 0.50$	CSI: $p = 0.25$
19	89	19.13	19.11	19.42	19.13	19.12	19.09	19.08	19.08	19.22	19.25	19.22
20	91	20.35	20.30	20.45	20.31	20.29	20.24	20.22	20.22	20.33	20.31	20.24
21	86	21.53	21.47	21.42	21.43	21.41	21.35	21.35	21.35	21.38	21.31	21.21
22	84	22.66	22.59	22.34	22.49	22.47	22.43	22.44	22.44	22.38	22.26	22.14
23	71	23.73	23.66	23.22	23.51	23.49	23.48	23.50	23.50	23.33	23.18	23.04
24	77	24.72	24.67	24.10	24.51	24.49	24.51	24.52	24.53	24.27	24.08	23.94
25	73	25.63	25.62	24.98	25.47	25.47	25.52	25.50	25.51	25.18	24.99	24.86
26	63	26.49	26.50	25.87	26.41	26.42	26.46	26.43	26.45	26.08	25.90	25.79
27	76	27.32	27.36	26.79	27.31	27.33	27.35	27.32	27.35	26.98	26.84	26.78
28	61	28.15	28.20	27.81	28.21	28.23	28.21	28.19	28.22	27.94	27.89	27.91
29	45	29.03	29.07	29.08	29.13	29.14	29.11	29.08	29.10	29.10	29.21	29.37
30	38	30.00	30.00	30.00	30.00	30.00	30.00	30.00	30.00	30.00	30.00	30.00

Note. AIRT = alternative item response theory, OSE = observed score equating, IRT = item response theory, TSE = true score equating, PCF = polynomial curve fitting, LSI = linear spline interpolation, CSI = cubic spline interpolation.

Table 1 contains the observed frequency, f_x, and the unrounded $\hat{\varphi}(x)$ values resulting from the three IRT equating methods. At a glance, there were large differences among $\hat{\varphi}(x)$ values when the raw scores of form X were low. For example, when x = 5, the rounded $\varphi(x)$ estimates for IRT-TSE and IRT-OSE were 5 and 4, respectively. Under AIRT-OSE, however, the rounded $\varphi(x)$ estimates appeared to be 1 or 2 for x = 5. But, for the high x scores such as 28 and 29, all the IRT equating methods produced very similar results in terms of rounded $\hat{\varphi}(x)$s.

As mentioned earlier, IRT-TSE was not able to estimate $\varphi(y)$s for a perfect score or an observed score of x less than the sum of guessing parameters. To circumvent the former problem, $\theta = 10$ was used in determining the equating equivalent corresponding to x = 30, $\hat{\varphi}(x = 30)$, on the test characteristic curve of form Y. To solve the latter problem, Kolen's (1981) ad hoc procedure was adopted. For instance, $\hat{\varphi}(x = 2)$ was calculated as $(5.35 / 5.21) \times 2 = 2.12$ when the sums of guessing parameters for forms X and Y were 5.21 and 5.35, respectively.

For the raw scores of form X ranging from zero through four, $\hat{\varphi}(x)$ values under AIRT-OSE were found to be zero, which resulted from the operational rule mentioned earlier. The $\hat{\theta}_x$ values for x = 0, 1, 2, 3 and 4 were smaller than the minimum $\hat{\theta}_y$ for y = 0. This case can happen when form X is easier than form Y for low ability levels. Because an examinee who got x = 0, 1, 2, 3, or 4 appeared to be less able than another examinee who got y = 0 in terms of EAP_{SS}, it is reasonable that $\hat{\varphi}(x)$ values for x = 0, 1, 2, 3, and 4 are zero.

Table 2

Job Skills Assessment Data: MSD, MAD, and RMSD Calculated with Two Sets of $\hat{\varphi}(x)$s

Difference Indices Equating Methods	MSD		MAD		RMSD	
	IRT-TSE	IRT-OSE	IRT-TSE	IRT-OSE	IRT-TSE	IRT-OSE
IRT-TSE	0	0.039	0	0.166	0	0.328
IRT-OSE	-0.039	0	0.166	0	0.328	0
AIRT-OSE						
PCF: 3rd Degree Poly.	-0.091	-0.052	0.576	0.425	0.739	0.484
PCF: 5th Degree Poly.	-0.029	0.009	0.272	0.145	0.599	0.305
PCF: 7th Degree Poly.	-0.034	0.005	0.284	0.139	0.593	0.298
PCF: 9th Degree Poly.	-0.037	0.001	0.293	0.138	0.587	0.295
LSI	-0.034	0.005	0.297	0.144	0.587	0.295
CSI: p = 1.00	-0.038	0.001	0.297	0.142	0.586	0.295
CSI: p = 0.75	-0.037	0.001	0.437	0.284	0.666	0.373
CSI: p = 0.50	-0.033	0.006	0.528	0.372	0.717	0.445
CSI: p = 0.25	-0.011	0.027	0.578	0.421	0.749	0.490

Note. MSD = mean signed difference, MAD = mean absolute difference, RMSD = root mean squared difference, IRT = item response theory, TSE = true score equating, OSE = observed score equating, AIRT = alternative item response theory, PCF = polynomial curve fitting, LSI = linear spline interpolation, CSI = cubic spline interpolation.

Table 2 presents the observed values of MSD, MAD and RMSD computed using the job skills assessment data. As shown in Table 2, the observed values of MSD, MAD, and RMSD for IRT-TSE versus IRT-OSE were 0.039, 0.166, and 0.328, respectively. Differences in the magnitudes of the three loss indices were found to be smaller for AIRT-OSE and IRT-OSE than for IRT-TSE and IRT-OSE except for the case of AIRT-OSE with third order degree polynomial curve fit. Also, it is noted that the observed differences in MAD values between IRT-OSE and AIRT-OSE were always smaller than those between IRT-TSE and AIRT-OSE. The same results were found for RMSD. In summary, the observed values for the three loss indices given in Table 2 indicate AIRT-OSE produced results closer to IRT-OSE than to IRT-TSE. Among the nine AIRT-OSE approaches, PCF with the ninth degree, LSI, and CSI with p = 1.00 provided the closest results to those for IRT-OSE in terms of RMSD (= 0.295).

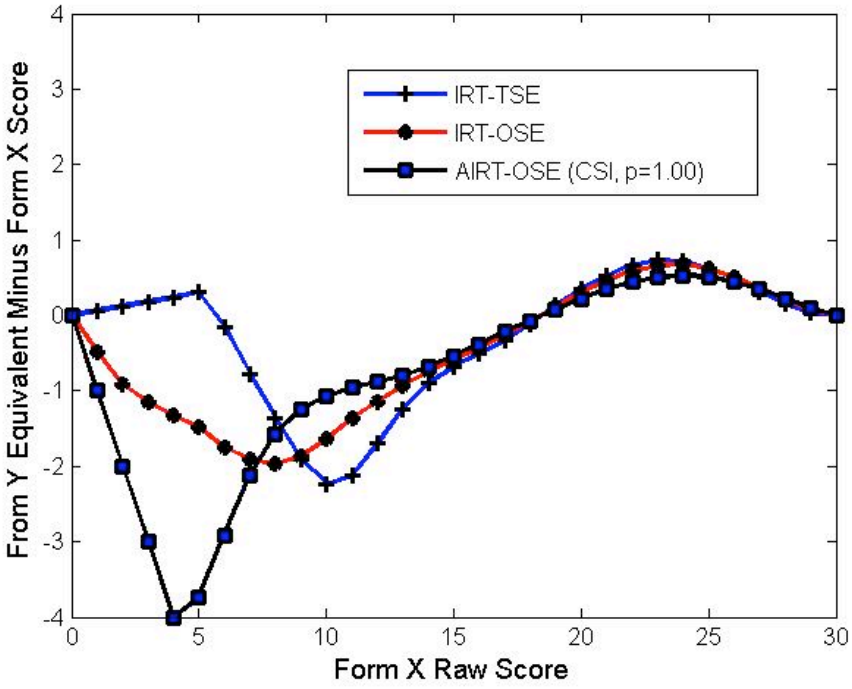

Figure 2. Job skills assessment data: Estimated equating patterns for the three IRT equating methods.

In Figure 2, the relationships between form X scores and their $\hat{\varphi}(x)$s under three IRT equating methods (IRT-TSE, IRT-OSE, and AIRT-OSE with CSI using p = 1.00) were plotted. Because for AIRT-OSE, the three cases (PCF with the ninth degree, LSI, and CSI with p = 1.00) yielded very similar results in terms of RMSD, only the last one was

considered in the following analysis. The three conversions exhibited in Figure 2 appeared noticeably different for low scores.

Between IRT-TSE and IRT-OSE, the largest difference was around x = 5 which is approximately equal to the sum of the guessing parameter estimates. This was coincidental with the finding of Kolen and Brennan (2004). It was interesting that for x = 0, 1, 2, 3, 4 and 5, IRT-TSE provided $\hat{\varphi}(x)$ values higher than x, whereas the $\hat{\varphi}(x)$s under both IRT-OSE and AIRT-OSE were less than the corresponding x values.

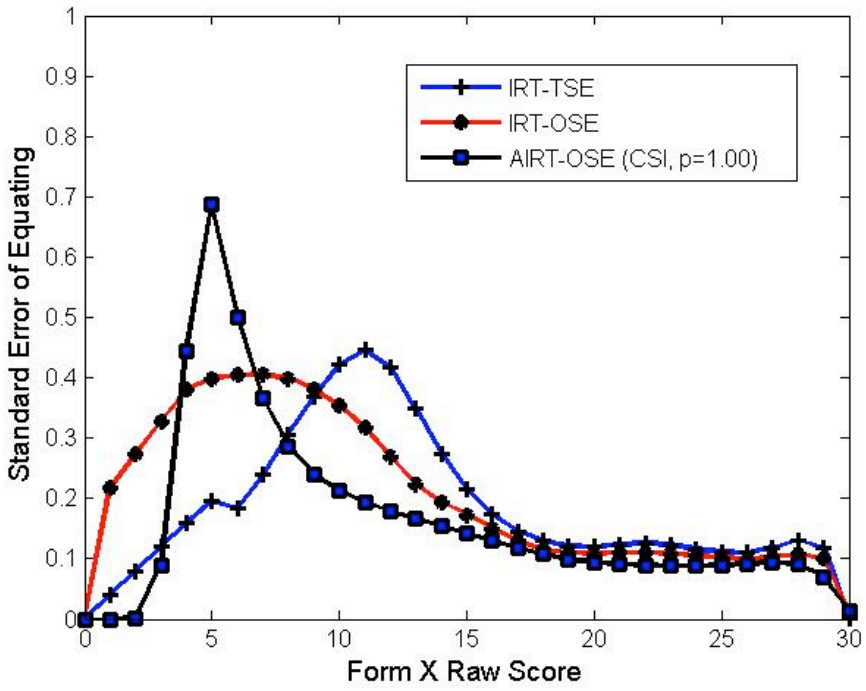

Figure 3. Job skills assessment data: Standard errors of equating.

Figure 3 compares the standard errors for the three IRT equating methods. For each IRT equating method, the bootstrap standard errors of equating were calculated according to Kolen and Brennan (2004). In the low x score range, the three equating methods showed large standard error differences. For x scores higher than 15, similar standard errors of equating around 0.10 were observed. The largest standard errors for AIRT-OSE, IRT-OSE, and IRT-TSE were, respectively, found to be 0.69 at x = 5, 0.41 at x = 7, and 0.45 at x = 11. For x = 8 or higher, the standard errors of AIRT-OSE consistently appeared to be the smallest in comparison to those of the two other equating methods.

Academic Achievement Assessment

This study used data from two forms of a mathematics achievement assessment with each consisting of 60 multiple-choice items. These two forms were, respectively, administered to two randomly equivalent groups of examinees. There were about 2,000 students in each group. The two forms did not share any common items, and each form's item parameters under the 3-PLM were estimated with BILOG (Mislevy & Bock, 1990). With the random equivalent group design, the item parameter estimates of the two forms were considered to be on common scale upon successful calibrations.

Table 3

Academic Achievement Assessment Data: MSD, MAD, and RMSD Calculated with Two Sets of $\hat{\varphi}(x)$ s.

Difference indices / Equating methods	MSD		MAD		RMSD	
	IRT-TSE	IRT-OSE	IRT-TSE	IRT-OSE	IRT-TSE	IRT-OSE
IRT-TSE	0	-0.005	0	0.024	0	0.079
IRT-OSE	0.005	0	0.024	0	0.079	0
AIRT-OSE						
PCF: 3rd degree poly.	-0.205	-0.210	0.659	0.661	0.742	0.733
PCF: 5th degree poly.	0.005	0.000	0.074	0.061	0.105	0.071
PCF: 7th degree poly.	0.006	0.001	0.053	0.036	0.095	0.048
PCF: 9th degree poly.	0.006	0.001	0.044	0.025	0.091	0.038
LSI	0.007	0.002	0.043	0.024	0.092	0.038
CSI: $p = 1.00$	0.006	0.001	0.043	0.025	0.092	0.039
CSI: $p = 0.75$	-0.045	-0.050	0.177	0.165	0.220	0.186
CSI: $p = 0.50$	-0.087	-0.092	0.362	0.357	0.416	0.398
CSI: $p = 0.25$	-0.122	-0.127	0.582	0.577	0.656	0.643

Note. MSD = mean signed difference, MAD = mean absolute difference, RMSD = root mean squared difference, IRT = item response theory, TSE = true score equating, OSE = observed score equating, AIRT = alternative item response theory, PCF = polynomial curve fitting, LSI = linear spline interpolation, CSI = cubic spline interpolation.

Table 3 presents the observed MSD, MAD, and RMSD computed using data from the two academic achievement assessment forms for the three IRT equating methods under study. The overall pattern of the three loss indices in Table 3 appears to be very similar to those for the job skills assessment in Table 2. Most of loss index values in Table 3 are much smaller than the corresponding values in Table 2. However, this might be explained by the difference in the number of items between the two assessments. In Table 3, AIRT-OSE

versus IRT-OSE, the observed magnitudes of the loss index in absolute values were found to be smaller than those for AIRT-OSE versus IRT-TSE in most of the cases. It can also be seen that the three AIRT-OSE cases including PCF with ninth degree polynomials, LSI, and CSI with p = 1.00 had very similar results for RMSD (e.g., 0.038 or 0.039). Also, the values for RMSD indicated that AIRT-OSE produced the closest equating results to those of IRT-OSE. To be consistent with the analyses of the previous real data set, only the case of AIRT-OSE with CSI (p = 1.00) was included in the following analysis.

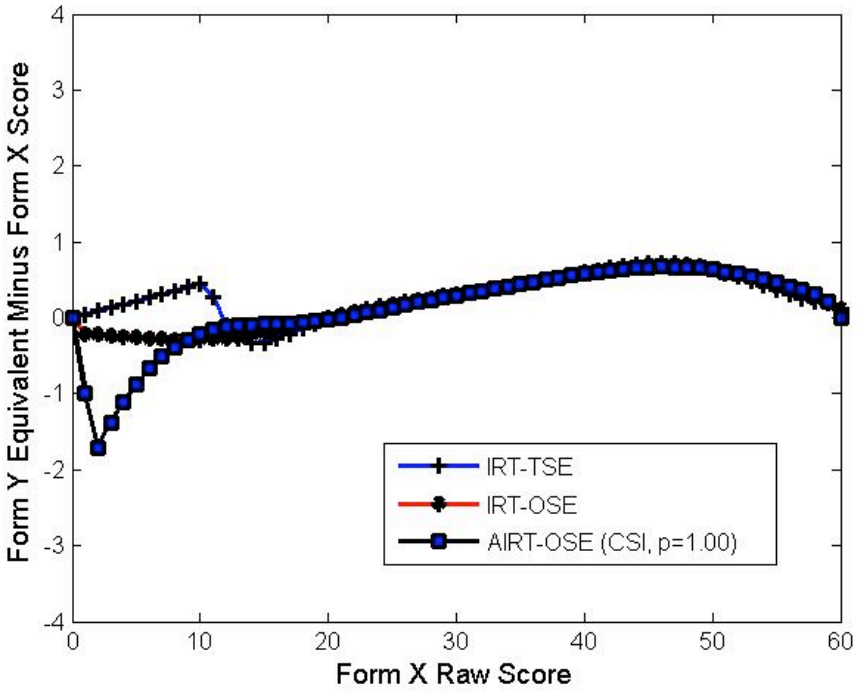

Figure 4. Academic achievement assessment data: Estimated equating patterns for the three IRT equating methods.

In Figure 4, the conversion relationships between new form scores and their $\hat{\varphi}(x)$s on the old form scale are plotted for each IRT equating method. For form X scores ranging from 1 through 12, three methods showed large differences in resulting $\hat{\varphi}(x)$s. For the other x scores, however, they produced very similar $\hat{\varphi}(x)$ values. For the low x score range, IRT-TSE had $\hat{\varphi}(x)$s higher than corresponding x scores; whereas both AIRT-OSE and IRT-OSE yielded $\hat{\varphi}(x)$ values lower than their corresponding x scores. The largest difference in $\hat{\varphi}(x)$ between IRT-TSE and IRT-OSE occurred at $x \approx 11$ which is approximately equal to the sum of the guessing parameter estimates. When x = 2, AIRT-OSE appeared to have the largest value of $\hat{\varphi}(x) - x$ among the three equating methods.

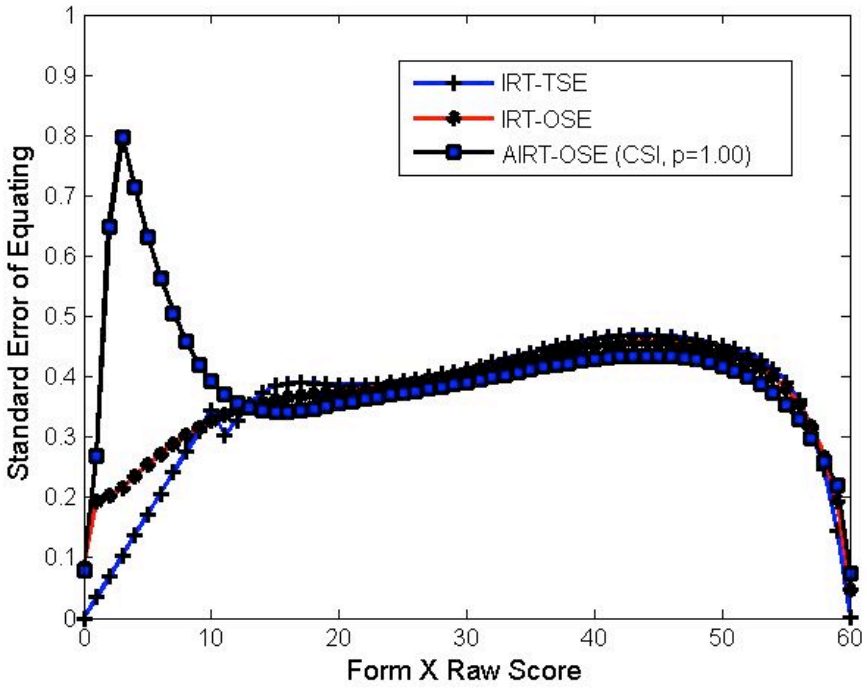

Figure 5. Academic achievement assessment data: Standard errors of equating.

Figure 5 exhibits the bootstrap standard errors of equating calculated for each IRT equating method applied to the academic achievement assessment data. The three IRT equating methods showed large standard error differences when the form X scores ranged between 0 and 11. For the x scores greater than 11, the standard errors of equating for the three methods appeared to be very similar and those for AIRT-OSE are the smallest in most cases.

Discussion

This report presents an alternative IRT observed score equating method that uses ability estimates based on summed scores. Exploiting the one-to-one relationship between EAP_{SS} and the observed number-correct score, this new method could perform observed score equating that technically resembles the IRT-TSE procedure. Because AIRT-OSE is based on observed scores, however, it was reasonable that the AIRT-OSE output was closer to that of IRT-OSE than that of IRT-TSE in terms of the three loss indices.

In this study, the AIRT-OSE method produced conversions and bootstrap standard errors for equating similar to those of the traditional IRT-TSE and IRT-OSE methods. In the low x score range, however, the three IRT equating methods tended to show very different equating performance. According to the range of the score scale that the test users are mostly

concerned about, a cautious application of the equating methods needs to be made as warned by Kolen and Brennan (2004). For example, if a cut-off score classifying examinees into master and non-master groups falls in a problematic score region, different equating methods could entail variant results.

As explained earlier, the implementation of AIRT-OSE method includes a 3-step procedure. In the third step, it is necessary to compute $\varphi(x)$ estimates through a curve fitting or interpolation strategy. To handle this problem, this study attempted three different approaches including PCF, LSI, and CSI. The analysis outcome from the two empirical data sets indicated that PCF with the ninth degree polynomial, LSI, and CSI with p = 1.00 produced similar equating results to those of IRT-OSE. In general, however, the spline approach tends to be preferred over the polynomial curve fitting because (a) PCF may be not flexible enough to fit various changes in real data points with low degree polynomials, and (b) PCF is apt to suffer from the problem of multicollinearity with high degree polynomials (Marsh & Cormier, 2001).

To better understand which method is the most appropriate in a given testing situation, however, further studies need to be conducted including a simulation study with different relevant factors (e.g., numbers of items, numbers of examinees, kinds of population distribution in Equation 2, dichotomous or polytomous IRT models, etc.). Also, it would be of interest to investigate the performance of AIRT-OSE under the Rasch model. Because a total test score is the sufficient statistic for an examinee's latent trait in the case, the AIRT-OSE approach can be applied without resorting to the use of EAP_{SS}.

REFERENCES

de Boor, C. (2001). *A practical guide to splines* (revised edition). New York: Springer-Verlag.

Efron, B. (1982). *The jackknife, the bootstrap, and other resampling plans*. Philadelphia: Society for Industrial and Applied Mathematics.

Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. Monographs on Statistics and Applied Probability, 57. New York: Chapman & Hall.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage Publications.

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.

Harris, D. J., & Crouse, J. D. (1993). A study of criteria used in equating. *Applied Measurement in Education, 6*, 195–240.

Kolen, M. J. (1981). Comparison of traditional and item response theory methods for equating tests. *Journal of Educational Measurement, 22*, 197–206.

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

Lord, F. M. (1977). Practical applications of item characteristic curve theory. *Journal of Educational Measurement, 14*, 117–135.

Lord, F. M. (1982). Item response theory and equating—A technical summary. In P. W. Holland & D. B. Rubin (Eds.), *Testing Equating* (pp. 141–161). New York: Academic.

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

Marsh, L. C., & Cormier, D. R. (2001). *Spline regression models* (Series: Quantitative Applications in the Social Sciences, Number 07-137). Thousand Oaks, CA: Sage Publications.

Mislevy, R. J., & Bock, R. D. (1990). *BILOG: Item analysis and test scoring with binary logistic models* [Computer Program]. Chicago: Scientific Software.

Reinsch, C. H. (1967). Smoothing by spline functions. *Numerische Mathematik, 10*, 177-183.

Schoenberg, I. J. (1964). Spline functions and the problem of graduation. *Proceedings of the National Academy of Science USA, 52*, 947–950.

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

The MathWorks. (2003). MATLAB (Version 6.5.1) [Computer Software]. Natick, MA: The MathWorks.

Thissen, D., Pommerich, M., Billeaud, K., & Williams, V. (1995). Item response theory for scores on tests including polytomous items with ordered responses. *Applied Psychological Measurement, 19*, 39–49.

Thissen, D., & Orlando, M. (2001). Item response theory for items scored in two categories. In D. Thissen and H. Wainer (Eds.), *Test scoring* (pp. 73–140). Mahwah, NJ: Lawrence Erlbaum Associates.