International Journal of Environmental Sciences
ISSN: 2229-7359
Vol. 11 No. 18s, 2025
https://www.theaspd.com/ijes.php

A Comparison Of Error Rates Of Classical Test Theory And Item Response Theory-Based Test Equating Methods Using Simulation Data

Şeyma Erbay Mermer

Bilecik Şeyh Edebali Univesity, Faculty of Health Sciences, Bilecik, Turkey. Email: sey.erbay@gmail.com
ORCID ID: 0000-0002-7747-9545

Abstract
In educational measurement and evaluation processes, the ability of different test forms to produce comparable scores is critical for obtaining fair and valid results. In this regard, the present study aims to comparatively examine test equating methods based on the Classical Test Theory (CTT) and the Item Response Theory (IRT), which are widely used in measurement applications, in terms of their error levels. Within this scope, using data generated through simulation, the methods of Linear Equating, Equipercentile Equating, Mean-Mean, Mean-Standard Deviation, Stocking-Lord, and Haebara were applied. Excel and R software were utilized for the analyses. In order to determine the amount of error between each raw score and its corresponding equated score, the Weighted Mean Squared Error (WMSE) criterion was used. These error values enabled a comparative evaluation of the accuracy levels of equating methods according to score ranges. The findings revealed that equating methods based on Item Response Theory produced lower error particularly at the extreme ends of the score distribution, whereas methods based on Classical Test Theory provided more consistent results at the mid-range. The study presents the error behaviors of test equating approaches through simulation-based comparisons and contributes to the theoretical discussions in the field of measurement.

Key Words: Test Equating, Item Response Theory, Classical Test Theory, Weighted Mean Squared Error (WMSE), Simulation.

Introduction

In educational measurement and evaluation processes, the comparison and interpretation of test scores are of great importance. When different test forms are administered, test equating methods are needed in order to make a fair and valid assessment of candidates' ability levels. Test equating enables the comparison of individuals' scores by expressing the scores obtained from different test forms on a common scale (Kolen & Brennan, 2014). In this context, test equating methods are generally classified into two main categories: observed score equating and true score equating. In the observed score equating method, the goal is to place scores on a common scale based on the score distributions of a specific group. This approach includes methods such as Linear Equating and Equipercentile Equating developed within the framework of Classical Test Theory (CTT) (von Davier, 2008). On the other hand, in true score equating, the aim is for an individual to obtain the same true score on different test forms depending on their ability level. These methods include various models developed within the framework of Item Response Theory (IRT), such as Mean-Mean, Mean-Standard Deviation, Stocking-Lord, and Haebara methods (Lord, 1980; Stocking & Lord, 1983). In true score equating, the true score is obtained by subtracting the error from the observed score, and its fundamental assumption is that individuals with the same ability level will have the same true score regardless of the test form they take (Lord, 1980). True score equating methods include both concurrent calibration (Lord, 1980) and separate calibration methods (Stocking & Lord, 1983).

In CTT-based equating methods, summary statistics at the test level, such as means and standard deviations of test scores, are used. Therefore, CTT-based equating is typically carried out using linear or equipercentile methods. These methods operate based on the relationship between total test scores and do not model item characteristics directly (Kolen & Brennan, 2014). On the other hand, IRT-based equating methods perform

analyses by taking into account item characteristics such as discrimination (a), difficulty (b), and guessing (c). In these methods, the relationship between individuals' ability levels and item responses is modeled probabilistically. IRT equating, especially with techniques such as Stocking-Lord and Haebara, allows different test forms to be expressed on a common scale through item parameters (Kim & Kolen, 2004; Lord, 1980). Within the scope of the present study, the equating methods used in the test equating process are addressed within the framework of CTT- and IRT-based approaches. Linear Equating and Equipercentile Equating methods were examined based on CTT; and under IRT, widely used methods for obtaining equated scores such as Mean-Mean, Mean-Standard Deviation, Stocking-Lord, and Haebara were included.

Equating Methods Based on Classical Test Theory (CTT)

Within the scope of the present study, the CTT-based equating methods, namely Linear Equating and Equipercentile Equating, were initially calculated using Excel.

Linear Equating

This method allows for differences in difficulty between two forms to vary across the scale scores. It assumes that scores corresponding to the same standard scores are equivalent (Angoff, 1984). Linear Equating is based on the division of the difference of scores from their own means by their standard deviations. Its distinction from mean equating lies in the standard deviation value in the following equation (Kolen & Brennan, 2014).

$$l_Y(X) = y = \frac{\sigma(Y)}{\sigma(X)}x + \left[\mu(Y) - \frac{\sigma(Y)}{\sigma(X)}\mu(X)\right]$$

X: Scores obtained from the X test (old form)
$\mu(X)$: Mean of the X test
Y: Scores obtained from the Y test (new form)
$\mu(Y)$: Mean of the Y test
$\sigma(X)$: Standard deviation of the X test
$\sigma(Y)$: Standard deviation of the Y test

Equipercentile Equating

In the equating function, if the distribution of scores from Form X, converted to the scale of Form Y, matches the distribution of scores from Form Y in the population, this constitutes an Equipercentile Equating function. This function is developed by identifying the scores from Form X that have the same percentile ranks as those in Form Y (Mutluer & Çakan, 2022).

Equipercentile Equating is based on identifying the scores on the tests to be equated that fall at the same percentile rank (Crocker & Algina, 1986). In this method, if the transformation of raw scores from the tests to be equated corresponds to the same percentile rank within the same population, these scores are considered equivalent.

$$e_y(X)=((P(X))/100-G(y_u^*-1))/(G(y_u^*)-G(y_u^*-1))+(y_u^*-0,5)$$

$e_y(X)$: Symmetric equating function that converts a score from Form X to Form Y
P(X): Percentile rank function for X
y_u^*: The smallest integer with a cumulative percentile greater than Q*
G(y): Discrete cumulative distribution of y

Equating Methods Based on Item Response Theory (IRT)

IRT-based equating enables meaningful comparison of test scores, particularly in cases where there are structural differences between test forms due to item parameters. In this regard, based on the item difficulty (b) and discrimination (a) parameters of the items included in different forms under the two-parameter logistic model, the aim is to express the scores that an individual would obtain from both forms on the same

scale.

At the core of the theory lies a mathematical model that describes how individuals at different ability levels respond to an item measuring a given trait. These equating methods are typically applied through common items (anchor items) or common individuals included in the test forms. In this way, the measurement results from different test forms become comparable through linear or nonlinear transformations (Baker, 1992; Crocker & Algina, 1986).

Baker (2001), Crocker and Algina (1986), Doğan and Tezbaşaran (2003), Embretson and Reise (2013), Hambleton and Swaminathan (2013), and Lord (1968) argued that using IRT is advantageous, as the examinee's ability can be identified with less measurement error and the item parameters exhibit invariance across different samples drawn from the same population.

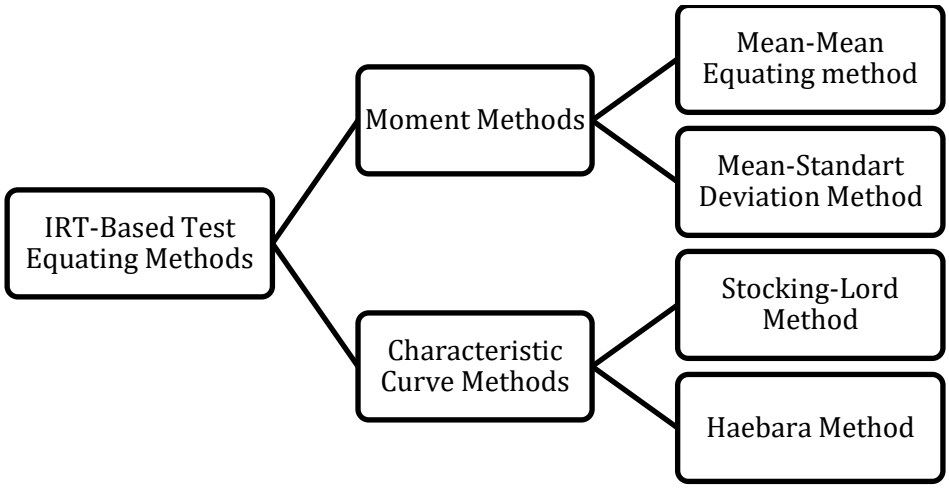

Figure 1. IRT-Based Test Equating Methods

As shown in Figure 1, IRT-based test equating methods are examined under two main categories: Moment Methods and Characteristic Curve Methods. Equating based on Moment Methods transforms scores using basic statistical values (mean, variance), while equating based on the Characteristic Curve Method performs more detailed transformations based on cumulative distribution functions by taking into account the entire structure of the score distribution.

IRT-based equating methods are of great importance in ensuring fair and comparable evaluations of measurement instruments. Compared to CTT-based equating approaches, these methods offer greater flexibility and allow for more precise analysis at the item level (Hambleton, Swaminathan & Rogers, 1991; Kolen & Brennan, 2014). Especially in high-stakes testing and longitudinal studies, reliable equating processes have become an indispensable component.

Mean-Mean (M-M) and Mean-Standard Deviation (M-S) Methods
Both methods are used to compute scaling constants. In the M-M method defined by Loyd and Hoover (1980), the average of the item discrimination parameters obtained from common items is used to estimate the A and B equating constants. The coefficients for this method are calculated as follows:

$$A = \frac{\mu(a_I)}{\mu(a_J)}$$

$$B = \mu(b_J) - A\mu(b_I)$$

$\mu(a_1)$: Discrimination parameter of Form I
$\mu(a_j)$: Discrimination parameter of Form J
$\mu(b_1)$: Difficulty parameter of Form I
$\mu(b_j)$: Difficulty parameter of Form J
A: Slope in the equating equation
B: Intercept (constant) in the equating equation
In the Mean-Mean (M-M) equating method, the scaling constant A is calculated first. This coefficient is obtained by dividing the average of the a parameters estimated from Form I by the average of the a parameters estimated from Form J. The A coefficient determined in this way is then used to calculate the second component of the equating process, the B constant. To calculate B, the product of the average b parameters from Form I and the A slope is subtracted from the average b parameters obtained from Form J.
Using these two coefficients, the individual item parameters from Form I are transformed to the scale of Form J. The M-M method offers a simple and applicable transformation approach, especially when the number of common items is sufficient and item parameters are reliably estimated (Kolen & Brennan, 2014).
In the Mean-Standard Deviation (M-S) method, defined by Marco (1977), the mean and standard deviation of the item difficulty parameters obtained from the common items are used to estimate the equating constants A and B. The equations used to estimate the A and B coefficients are provided below (Kolen & Brennan, 2014; Loyd & Hoover, 1980):

$$A = \frac{\sigma(b_J)}{\sigma(b_I)}$$

$$B = \mu(b_J) - A\mu(b_I)$$

$\sigma(b_j)$: Standard deviation of the difficulty parameters of Form J
$\sigma(b_i)$: Standard deviation of the difficulty parameters of Form I
$\mu(b_j)$: Mean of the difficulty parameters of Form J
$\mu(b_i)$: Mean of the difficulty parameters of Form I
A: Slope in the equating equation
B: Intercept (constant) in the equating equation
In the M-S equating method, first, the scaling coefficient denoted as term A is calculated. The A term is obtained by dividing the standard deviation of the a parameter computed after applying Form I by the standard deviation of the a parameter calculated from Form J. The A term obtained through this equating method is then used in the second equation. To calculate the constant B, the result obtained by multiplying the mean of the b parameters in Form I by the slope value A is subtracted from the mean of the b parameters obtained from Form J. Thus, the constant B is also determined.
Although the moment methods estimation is based on items with similar item characteristic curves but differing difficulty parameters, the mean-standard deviation method will be significantly affected due to differences in the b parameter estimations. The reason for this is that these methods do not consider all item parameter estimations simultaneously during scale transformation.

Haebara and Stocking and Lord (SL) Methods

Haebara Method: For a given ability level, the difference between item characteristic curves refers to the sum of the squared differences between the item characteristic curves for each item. The mathematical representation of this method is as follows:

$$Hdiff(\theta_i) = \sum_{j=V} \left[p_{ij}(\theta_{Ji};\ \hat{a}_{Jj}, \hat{b}_{Jj}, \hat{c}_{Jj}) - p_{ij}\left(\theta_{Ji};\ \frac{\hat{a}_{Ij}}{A}, A\hat{b}_{Ij} + B, \hat{c}_{Ij}\right)\right]^2$$

A = Slope in the equating equation

B = Intercept in the equating equation
$p_{ij}(\theta_{ji}; \hat{a}_{Jj}, \hat{b}_{Jj}, \hat{c}_{Jj})$: Item characteristic curve
$p_{ij}\left(\theta_{ji}; \frac{\hat{a}_{Ij}}{A}, A\hat{b}_{Ij} + B, \hat{c}_{Ij}\right)$: Estimated item characteristic curve
Hdiff(θ_i): Sum of squared differences between the item characteristic curves for each item in the test at a given ability level θ_i

The objective is to minimize the difference between the Item Characteristic Curves (ICCs) of the common items across the two test forms.

Stocking and Lord (SL) Method: Unlike the Haebara method, the SL method uses the squared difference of the sums. Before squaring, it takes the sum of the item characteristic curves for each set of items.

$$SLdiff(\theta_i) = \left[\sum_{j=V} p_{ij}(\theta_{Ji};\ \hat{a}_{Jj}, \hat{b}_{Jj}, \hat{c}_{Jj}) - \sum_{j=V} p_{ij}\left(\theta_{Ji};\ \frac{\hat{a}_{Ij}}{A}, A\hat{b}_{Ij} + B, \hat{c}_{Ij}\right)\right]^2$$

A = Slope in the equating equation
B = Intercept in the equating equation
$p_{ij}(\theta_{ji}; \hat{a}_{Jj}, \hat{b}_{Jj}, \hat{c}_{Jj})$: Item characteristic curve
$p_{ij}\left(\theta_{ji}; \frac{\hat{a}_{Ij}}{A}, A\hat{b}_{Ij} + B, \hat{c}_{Ij}\right)$: Estimated item characteristic curve
SLdiff(θ_i): The squared difference between the total sums of item characteristic curves across the test for each ability level θ_i
This method determines the scale equating coefficients (A and B) by minimizing the difference between the Test Characteristic Curves (TCC). In both item characteristic curve-based methods, the coefficients A and B, which minimize the resulting functions, can be estimated using the coefficients obtained from moment methods as initial values for the iteration process.
Models developed based on IRT allow for more sensitive and individually oriented interpretation of measurement results in the field of test equating. Thanks to these features, they are more widely preferred compared to classical approaches (Kolen & Brennan, 2014). However, the applicability and equating accuracy of these methods are influenced by various factors, such as test length, the number of common items, group equivalence, and sample size. In particular, when equating is performed with a limited number of common items in non-equivalent groups, increased error rates necessitate a re-evaluation of the validity and reliability of these methods (Kim & Kolen, 2004).

Error in Test Equating
The standard error in test equating is the standard deviation of score similarities across repeated equating processes, and this error is divided into random and systematic components (Mutluer & Çakan, 2023). Random error arises from the estimation process of equating parameters and is generally expressed as the standard error of equating (Gök & Kelecioğlu, 2014). This error can be reduced by using a large sample and an appropriate equating design. Systematic error, on the other hand, occurs when the assumptions or conditions of equating are violated. There is no standard method for determining the amount of such error (Kolen & Brennan, 2014).
Lehman and Bailey (1968) stated that simulation studies can be conducted with ease in cases where empirical research is not feasible or would be too costly. Similarly, Oh (2000) emphasized that the amount of data typically required for analysis can often only be obtained through simulation, which is why simulated data are frequently used in equating studies. This makes simulation an attractive alternative for exploring methodological questions and evaluating the performance of statistical techniques under controlled conditions. In Turkey, although the constructs assessed in large-scale and recurring exams (such as YKS, KPSS, and LGS) remain consistent, a new test—assumed to be parallel to previous versions—is administered

each year due to test security concerns. However, none of these test forms include common items linking them to previous years or to other forms. Given the difficulty in ensuring the equivalence of individuals who take the test in different years or retake the same test, implementing an equivalent-groups design becomes problematic. Moreover, the absence of common items in national exams makes it difficult to perform equating using the common-item design. Therefore, in this study, the performance of estimation methods based on CTT and IRT was compared through a simulation study.

Within the scope of the present study, it is aimed to comparatively examine the test equating methods based on CTT and IRT in terms of error levels. In this context, methods including Linear Equating, Equipercentile Equating, Mean-Mean, Mean-Standard Deviation, Stocking-Lord, and Haebara were applied using data generated through simulation.

Method

Research Design

The present study falls within the scope of fundamental research. Fundamental research is typically conducted to understand and explain the processes underlying a theory or hypothesis (Fraenkel, Wallen, & Hyun, 2012). In the context of this study, it was aimed to compare the Classical Test Theory (linear equating and equipercentile equating) and Item Response Theory true score equating methods (M-M, M-S, SL, and Haebara) based on equating errors, within a simulation design involving common items in nonequivalent groups. The primary objective of this study is to determine which method yields the lowest error rate and to contribute to theoretical research.

Simulation studies are classified into two categories: deterministic and probabilistic simulation studies (Çörtük & Sinan, 2023). In deterministic simulations, all conditions are assumed to be fixed, whereas in probabilistic simulations, at least one condition is treated as a variable involving randomness (Cohen, Manion, & Morrison, 2007). Within this framework, the present study is characterized as a probabilistic simulation study in which at least one condition is randomly varied.

Simulation Conditions

The simulated data used in the study were generated using the freely accessible WinGen3 software developed by Han and Hambleton (2007). The conditions that were kept constant throughout the data generation process are presented in Table 1**.**

Table 1. Conditions Held Constant

Parameter	**Minimum Value**	**Maximum Value**
a parameter	0.25	1.10
b parameter	-3	3
Ability parameter	-3	3
Number of replications	20	–
Number of options	2	–
Number of factors	1	–
Test length	20	–
Internal anchor item	5	–
Model	2PLM	–

In the present study, 20-item tests were employed, with each test form structured to include 5 common items. A common-item design for nonequivalent groups was adopted based on dichotomous simulation data, with the sample size for each group set at 1,000 participants. In order to enhance the generalizability of the simulation results, replications (iterations) were conducted. In IRT-based studies, Harwell, Stone, Hsu, and Kirisci (1996) suggested that 25 replications are sufficient, while Kolen and Brennan (2014) stated that between 20 and 100 replications are adequate. Accordingly, 20 replications were conducted for each sub-condition in the current study.

A Monte Carlo simulation study was performed to identify the conditions under which the equating methods yield more accurate results. Item responses were generated based on dichotomous item responses using the two-parameter logistic model (2PL). Data analyses were conducted using the R software and Microsoft Excel. For IRT analyses in R, the "mirt", "stats4", and "lattice" packages were utilized—specifically for examining equating error and standard error effects.

In nonequivalent groups, a common-item design is often preferred in situations where multiple test forms are administered on the same day to maintain test security (Tsai, 1997). Since this design is applied to nonequivalent groups, it is referred to as a nonequivalent groups with anchor test (NEAT) design (von Davier et al., 2004). In this context, each group receives only one form administered a single time (Kolen, 1988). The NEAT design is presented in Table 2.

Table 2. Common-Item Design in Nonequivalent Groups

Group	Form1	Form2	Common Test
X	+		+
Y		+	+

One of the test forms and a set of common items are administered to nonequivalent groups, and the equating process is carried out based on these common items. In this design, the common items should adequately represent the test, and they must be placed in the same position in both forms, as the location of the items can affect their difficulty (Kolen, 1988; Kolen & Brennan, 2004; Tsai, 1997). Based on the placement of the items within the test form, the anchor test can be classified as either internal anchor or external anchor (Kolen & Brennan, 2014). If the scores obtained from the anchor items are included in the individual's total test score, it is referred to as an internal anchor; if they are not included, it is considered an external anchor. In the external anchor design, the anchor items are administered as a separate test form, while in the internal anchor design, they are embedded within the test (Cook & Eignor, 1991; Kolen, 1988; Zhu, 1998). In the present study, the anchor items were embedded within the test forms and implemented based on the internal anchor design.

Findings

In the present study, among the true score equating methods, characteristic curve methods (Stocking-Lord and Haebara) and moment methods (Mean-Mean and Mean-Sigma) were examined, alongside observed score equating methods including linear and equipercentile equating.

During the equating process, the Y test form was designated as the reference form, while the X form was the one to be equated. The sample size was determined to be 1000 individuals for each test form. In the study, the internal anchor item design was preferred as the common-item type, with anchor items comprising 25% of the total test items.

Table 3. Frequency and Percentile Values of Form Y and Form X

Score	Frequency		Percentile	
	Form Y	Form X	Form Y	Form X
0	0	0	0	0,0
1	1	2	0,1	0,2
2	5	15	0,5	1,5
3	8	32	0,8	3,2
4	33	42	3,3	4,2

5	41	87	4,1	8,7
6	64	96	6,4	9,6
7	90	96	9,0	9,6
8	119	92	11,9	9,2
9	112	104	11,2	10,4
10	103	93	10,3	9,3
11	91	78	9,1	7,8
12	94	71	9,4	7,1
13	71	55	7,1	5,5
14	54	46	5,4	4,6
15	42	29	4,2	2,9
16	31	20	3,1	2,0
17	21	20	2,1	2,0
18	12	12	1,2	1,2
19	4	6	0,4	0,6
20	4	4	0,4	0,4
Total	1000	1000	100	100

The table above presents the frequency and percentile values of the scores for Form Y and Form X. Within the scope of equating, the number of individuals was determined as 1000 for each form. It is observed that the score distribution is concentrated around scores of 8, 9, and 10.

Table 4. Descriptive Statistics for Form Y and Form X

	Mean	ss	Skewness	Kurtosis
FormY	10,029	3,465	0,253	-0,335
FormX	9,144	3,757	0,389	-0,327

According to Table 4, since the skewness and kurtosis coefficients fall within the range of -1.5 to +1.5, it can be interpreted that the individuals to whom Form Y and Form X were administered are homogeneously distributed (Tabachnick and Fidell, 2013). Crocker and Algina (1986) stated that, in the common-item test design, for IRT true score equating methods, the common test and the total test must measure the same trait/ability, whereas such a requirement is not necessary for traditional equating methods such as linear and equipercentile equating.

Table 5. Scores Obtained Through Linear and Equipercentile Equating Methods Corresponding to the i-th Raw Score on Form X

X Form Raw Scores	Linear Equating	Equipercentile Equating
1	2,518	1,500
2	3,440	3,200
3	4,362	5,875
4	5,285	5,197
5	6,207	6,634
6	7,129	7,656
7	8,052	8,389
8	8,974	8,962
9	9,896	9,866

10	10,818	10,854
11	11,741	11,841
12	12,663	12,622
13	13,585	13,549
14	14,507	14,500
15	15,430	15,393
16	16,532	16,145
17	17,274	16,929
18	18,197	17,833
19	19,199	18,750
20	20,041	20,000

Overall, the scores obtained through the linear equating method yield higher or lower results than those derived from the equipercentile equating method within certain score intervals. This discrepancy stems from the underlying assumptions of the equating types. In the lower score range (e.g., between 2 and 6), the equipercentile equating method produced lower scores compared to linear equating. This is associated with the equipercentile method's ability to more sensitively reflect differences at the lower end of the distribution. In the mid-score range (7–15), the difference between the two methods gradually decreases, indicating that both methods produce more similar results within this range. At higher score levels (16 and above), the equated scores converge significantly, with the equipercentile method yielding lower scores in some instances and higher scores in others. These differences can be attributed to the assumption of a constant relationship across all points of the distribution in linear equating, whereas the equipercentile method establishes a more flexible relationship by matching percentiles to corresponding scores.

To determine the equipercentile equivalent of a specific score on Form X, the score on Form Y that corresponds to the same percentile rank is identified. The fundamental assumption of this process is that individuals with the same level of achievement should occupy the same percentile rank across different test forms. For example, if an individual scores 65 on Form X and this corresponds to the 70th percentile, the equivalent score on the reference test form, Form Y, is the score that corresponds to the 70th percentile on that form. In this way, the achievement levels of individuals can be compared independently of the test form (Kolen & Brennan, 2014).

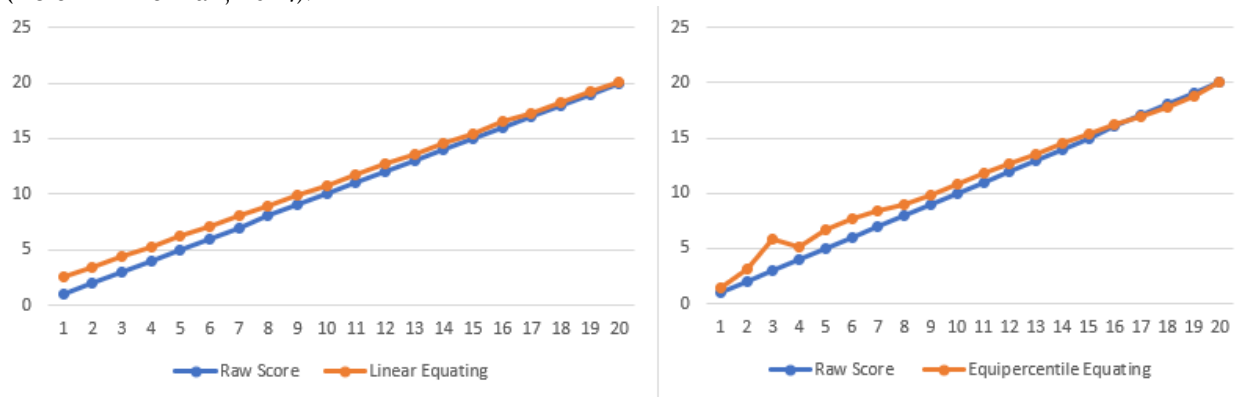

Figure 2. Equating with Linear and Equipercentile Equating Methods

When Figure 2 is examined, it can be observed how much the equated scores deviate from the raw scores. A linear relationship is evident between the equated and raw scores, and it is observed that greater deviations occur in lower scores when using the equipercentile equating method.

Equating Methods Based on IRT

Within the scope of IRT, item characteristic curve methods (SL and Haebara) and moment methods (M-M and M-S), which are estimation approaches different from true score equating methods, were examined.

Table 6. Item Difficulty and Discrimination Parameters for Form X and Form Y

	FormX		FormY	
	b Parameter	a Parameter	b Parameter	a Parameter
S1	-0,83	0,85	-0,85	0,62
S2	-0,77	0,99	-0,77	0,91
S3	-0,22	0,25	-0,22	0,43
S4	0,02	1,15	0,02	0,92
S5	0,21	0,61	0,21	0,55
S6	-0,52	1,12	0,24	0,81
S7	-0,40	0,35	-0,64	0,57
S8	-0,75	0,66	0,22	1,10
S9	0,14	0,71	0,66	0,94
S10	-1,06	0,74	-0,64	0,54
S11	0,38	0,50	0,00	1,00
S12	0,38	1,71	-0,38	0,32
S13	0,38	1,22	0,72	0,83
S14	0,18	0,77	0,19	0,25
S15	0,13	0,26	1,26	0,81
S16	-0,30	0,91	1,13	0,88
S17	0,46	0,69	-0,56	0,53
S18	-0,51	0,87	-0,39	0,68
S19	-0,81	0,45	-0,66	0,46
S20	0,11	0,77	1,14	0,69

The table above presents the item difficulty (b) and discrimination (a) parameters for Form X and Form Y. Items with highly similar difficulty values have been designated as anchor items, labeled as s1, s2, s3, s4, and s5.

Table 7. Test Forms Equated Using IRT-Based Equating Methods

Raw Scores	M-M	M-S	SL	Haebara
1	1,35	1,17	1,20	1,50
2	2,42	2,24	2,18	2,54
3	3,49	3,33	3,28	3,58
4	4,54	4,43	4,37	4,63
5	5,59	5,52	5,50	5,68
6	6,65	6,62	6,58	6,72
7	7,70	7,71	7,60	7,75
8	8,75	8,81	8,50	8,77

9	9,79	9,91	9,60	9,79
10	10,82	10,99	10,70	10,79
11	11,84	12,05	11,70	11,77
12	12,84	13,09	12,60	12,75
13	13,82	14,09	13,50	13,70
14	14,77	15,07	14,61	14,63
15	15,70	16,01	15,40	15,54
16	16,59	16,91	16,40	16,44
17	17,46	17,78	17,31	17,35
18	18,36	18,57	18,22	18,26
19	19,26	19,40	19,14	19,17
20	20,18	20,29	20,09	20,10

Table 7 demonstrates how different equating methods based on IRT may vary, particularly at low and moderate score levels. As raw scores increase, the equated scores obtained through all methods also increase, indicating that the transformation proceeds in a monotonic and meaningful manner. Although all methods fundamentally follow the same transformation trend, the differences among methods become especially apparent at low and moderate score ranges. Haebara positions individuals with low achievement levels somewhat more "advantageously" on the reference form. SL performs a more cautious equating, keeping low scores relatively lower. This reduces the risk of score inflation but may potentially underrepresent the abilities of some individuals. M-M and M–S are simpler and linear methods; however, as they do not take item characteristics into account, they are not recommended particularly in the presence of heterogeneous item structures.

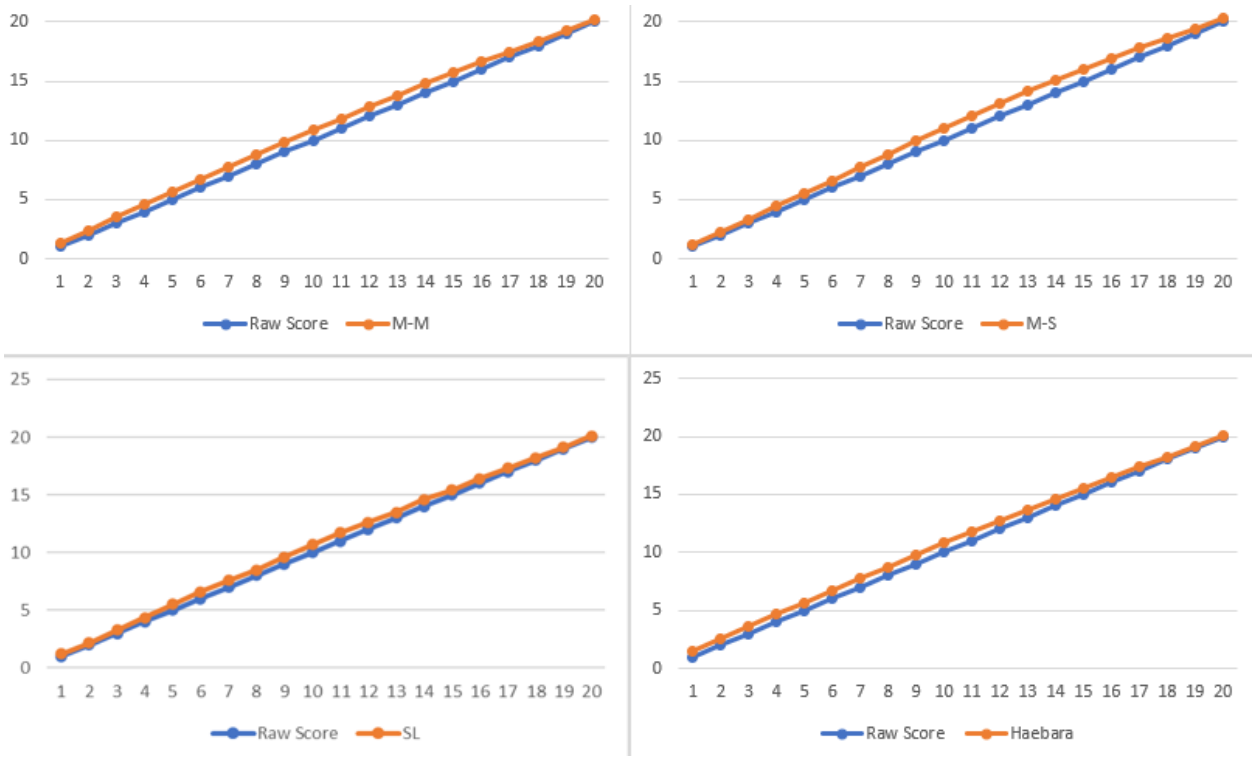

Figure 3. Equating with M-M, M–S, SL, and Haebara Methods

Figure 3 clearly illustrates how the choice of equating method results in differences across various score ranges. The graphs show that the M-M method provides a linear transformation, meaning it converts raw

scores to the reference form scores with a constant increment. The M–S method behaves similarly to M-M at lower score levels, but becomes more "generous" at higher scores, producing relatively higher values in the reference form. The Stocking–Lord (SL) curve starts close to the raw scores but increases less steeply as scores rise, indicating a more balanced and controlled equating. The Haebara method, on the other hand, provides the highest equating values at lower score levels, distinguishing it from the others.

Table 8. Comparison of Equating Errors between CTT and IRT Methods

	CTT		IRT			
Method	Linear	Equipercentile Equating	M-M	M-S	SL	Haebara
Equating Error	0,0614	0,0823	0,0496	0,0672	0,0422	0,0428

Table 8 presents a comparative summary of equating errors for different test equating methods based on Classical Test Theory (CTT) and Item Response Theory (IRT). The differences in equating errors arise from the conceptual frameworks of the methods. One such metric is the Weighted Mean Squared Error (WMSE). To determine the magnitude of error in the scores obtained through equating methods, each raw score and its corresponding equated score are compared using WMSE (Skaggs & Lissitz, 1986; Kelecioğlu, 1994; Şahhüseyinoğlu, 2005; Bozdağ & Kan, 2010).

The equating accuracy of each method is assessed based on its error value; lower error implies higher accuracy and reliability. According to Table 7, IRT-based methods perform equating with lower errors compared to those based on CTT. Among these, the Stocking–Lord and Haebara methods stand out as preferred options in academic research and operational test equating applications. CTT-based methods, on the other hand, may be used when the data structure is simple, test forms are highly similar, or when IRT cannot be applied.

Discussion and Conclusion

As the result of this study it is seen that methods have different results in different methods. The present study aimed to compare equating methods developed based on CTT and IRT. Within this scope, the linear equating and equipercentile equating methods from CTT, and the M-M, M-S, SL and Haebara methods from IRT were employed. Overall, IRT-based equating methods yielded more reliable results compared to those based on CTT. Consistent with the literature, IRT-based methods offer higher accuracy, better parametric control, and minimize distortions at the score extremes (Kim & Kolen, 2006).

Findings from the CTT framework revealed that the equipercentile equating method produced more sensitive results at the distributional extremes—particularly at lower score levels—than linear equating. Therefore, if score distributions deviate from normality or there are structural differences between test forms, the equipercentile method may be a more suitable choice. Among the IRT methods, Stocking–Lord and Haebara methods stood out in terms of graphical fit and lower error values. While CTT methods may be practical when the test forms are structurally similar, the sample sensitivity of the equipercentile method should be taken into account.

When equating errors are examined, IRT-based methods—especially Stocking–Lord and Haebara—demonstrated more accurate and stable equating. Based on these results, IRT methods are recommended for fair comparisons in education and for making valid decisions (Kolen & Brennan, 2014). The distinction between the Haebara and Stocking–Lord methods becomes particularly important when decisions focus on lower performance levels (e.g., failure thresholds). For high-stakes decisions based on high scores (e.g., scholarships, rankings), the M–S method may result in slightly inflated outcomes. For mid-range scores, all methods generally yield similar results. To determine the equating method according to the conditions taken into consideration in this study is significant. Due to the differences in the methods, the results do not show enough stability to prefer one method to the other and it can be said that there is no effective single method for every condition.

Recommendations

This study is limited to simulation data, dichotomously scored item responses, and unidimensionality. Future research could investigate the effect of multidimensionality on test equating using both concurrent and separate calibration methods for polytomous and multidimensional tests. In the equating of mixed-format tests, concurrent and separate calibration approaches may be compared. Although simulation data provide researchers with the opportunity to study multiple factors simultaneously (Harris & Crouse, 1993), they may not fully reflect real testing conditions. Therefore, it would be beneficial to test all the conditions considered and recommended in this study using real data sets.

REFERENCES

1. Angoff, W. H. (1982). Summary and derivation of equating methods used at ETS. In P.W. Holland ve D. B. Rubin (Ed). Test Equating. New York: Academic Press.
2. Baker, F. B. (1992). Equating tests under the graded response model. Applied Psychological Measurement, 16(1), 87-96.
3. Bozdağ, S., Kan, A. (2010). Şans Başarısının Test Eşitlemeye Etkisi. Hacettepe University Journal of Education, 39: 91-108.
4. Cook, L. L. & Eignor, D. R. (1991). An NCME instructional module on IRT equating methods. Educational measurement: Issues and Practice. 10 (3), 37-45.
5. Cohen, L., Manion, L., & Morrison, K. (2007). Internet-based research and computer usage. Research Methods in Education (6. Edition, s. 226-252). New York: Routledge.
6. Çörtük, M. & Sinan, A. (2023). Çok kategorili puanlanan maddelerden oluşan testlerde klasik test kuramı ve madde tepki kuramı'na dayalı test eşitleme yöntemlerinin karşılaştırılması. Turkish Academic Research Review, 8 (4), 1429-1439.
7. Crocker, L. & Algina, J. (1986). Introduction to classical and modern test theory. New York: Holt, Rinehart & Winston.
8. Demirus, K.B. (2015). Ortak maddelerin değişen madde fonksiyonu gösterip göstermemesi durumunda test eşitlemeye etkisinin farklı yöntemlerle incelenmesi (Unpublished doctoral dissertation). Hacettepe University, Graduate School of Educational Sciences, Ankara.
9. Doğan, N., & Tezbaşaran A. A. (2003). Klasik test kuramı ve örtük özellikler kuramının örneklemler bağlamında karşılaştırılması. Hacettepe University Journal of Education, 25(25), 58-67.
10. Embretson, S. E., & Reise, S. P. (2013). Item response theory. New Jersey: Psychology Press.
11. Fraenkel, J. R., Wallen, N. E., & Hyun, H. H. (2012). The nature of research. How to design and evaluate research in education (8. Edition). New York,NY: McGraw-Hill Education.
12. Gök, B., & Kelecioğlu, H. (2014). Denk olmayan gruplarda ortak madde deseni kullanılarak madde tepki kuramına dayalı eşitleme yöntemlerinin karşılaştırılması. Mersin University Journal of Education, 10(1), 120–136.
13. Haebara, T. (1980). Equating logistic ability scales by a weighted least squares method. Japanese Psychological Research, 22, 144-149.
14. Han, K. T., & Hambleton, R. K. (2007). User's manual: WinGen (642). Retrieved from Amherst, MA:University of Massachusetts.
15. Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). Fundamentals of Item Response Theory. Sage.
16. Hambleton, R. K., & Swaminathan, H. (2013). Item response theory: Principles and applications. New York: Springer Science & Business Media.
17. Harris, D. J., & Crouse, J. D. (1993). A study of criteria used in equating. Applied Measurement in Education, 6(4), 195–240.
18. Harwell, M., Stone, C. A., Hsu, T., & Kirisci, L. (1996). Monte Carlo studies in Item Response Theory. Applied Psychological Measurement, 20(2), 101-125.
19. Kelecioğlu, H. (1994). Öğrenci Seçme Sınavı Puanlarının Eşitlenmesi Üzerine Bir Çalışma (Doctoral Thesis). Hacettepe University, Graduate School of Social Sciences. Ankara.
20. Kim, S., & Kolen, M. J. (2004). STUIRT: A computer program for scale transformation under unidimensional item response theory models. Applied Psychological Measurement, 28(1), 64–82.
21. Kim, S., & Kolen, M. J. (2006). Comparison of traditional and item response theory methods for equating tests. Journal of Educational Measurement, 18(1), 1–11.
22. Kolen, M. J. (1988). An NCME instructional module on traditional equating methodology. Educational Measurement: Issues and Practice, 7, 29-36.
23. Kolen, M. J., & Brennan, R. L. (2014). Test Equating, Scaling, and Linking: Methods and Practices (3rd ed.). New

York: Springer.
24. Lehman, R. S., & Bailey, D. E. (1968). Digital computing: Fortran IV and its applications in behavioral science. New York: John Wiley.
25. Lord, F. M., & Novick M. R. (1968). Statistical theories of mental test scores. Reading, Mass: Addison-Wesley.
26. Lord, F. M. (1980). Applications of Item Response Theory to Practical Testing Problems. Hillsdale, NJ: Lawrence Erlbaum Associates.
27. Loyd, B. H., & Hoover, H. (1980). Vertical equating using the Rasch model. Journal of Educational Measurement, 17(3), 179-193.
28. Marco, G. L. (1977). Item characteristic curve solutions to three intractable testing problems. Journal of Educational Measurement, 14, 139-160.
29. Mutluer, C. & Çakan, M. (2023). Comparison of Test Equating Methods Based On Classical Test Theory and Item Response Theory. Journal of Uludağ University Faculty of Education, 36(3), 866–906.
30. Oh, S. (2000). Comparison of traditional and item response theory equating using arm and shoulder girdle muscular strength and endurance tests. (Doctorate thesis) University of Georgia, Athens, Georgia.
31. Skaggs, G., & Lissitz, R. W. (1986). IRT test equating: Relevant issues and a review of recent research. Review of Educational Research, 56(4), 495-529.
32. Stocking, M. L. & Lord, F. M. (1983). Developing a Common Metric in Item Response Theory. Applied Psychological Measurement, 7(2), 201-210.
33. Şahhüseyinoğlu, D. (2006). İngilizce yeterlik sınavı puanlarının üç farklı eşitleme yöntemine göre karşılaştırılması. Hacettepe University Journal of Education, (31), 115-125.
34. Tabachnick, B.G. & Fidell, L. S. (2013). Using multivariate statistics, 6th edition. Boston:Pearson.
35. Tsai, T. H. (1997). Estimating minumum sample sizes in random groups equating. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago.
36. von Davier, A. A., & Han, N. (2004). Population invariance and linear equating for the non-equivalent groups design (NEAT). ETS Research Report No. RR-04-47. Educational Testing Service.
37. von Davier, A.A. (2008). New results on the linear equating methods for the non- equivalentgroups design. Journal of Educational and Behavioral Statistics, 33(2), 186-203.
38. Zhu, W. (1998). Test equating: What, why, how?. Research Quarterly for Exercise and Sport, 69(1), 11-23.