Kastamonu Education Journal, 2025, Vol. 33, No:3, 658-671
doi: 10.24106/kefdergi.1750267

| Research Article / Araştırma Makalesi |

Examination of Scale Transformation and Test Equating Methods in Testlet Based Tests

Madde Takımı İçeren Testlerde Ölçek Dönüştürme ve Test Eşitleme Yöntemlerinin İncelenmesi[1]

Harun DİLEK[2], Kübra ATALAY KABASAKAL[3], Sebahat GÖREN[4]

Keywords

1. Testlet
2. Testlet effect
3. Test equating
4. Testlet Response Theory

Anahtar Kelimeler

1. Madde takımı
2. Madde takımı etkisi
3. Test eşitleme
4. Madde Takımı Tepki Kuramı

Received/Başvuru Tarihi
25.12.2024

Accepted / Kabul Tarihi
25.07.2025

Abstract

Purpose: This study examines the test equating performance under various item response theory models and sample size conditions in testlet based tests.

Design/Methodology/Approach: Utilizing data from the eTIMSS 2019 science test, the study compares scale transformation methods and test equating results under Unidimensional Item Response Theory (UIRT), Testlet Response Theory (TRT) and bifactor models with varying sample sizes. Scale transformation methods, including the mean-sigma and Stocking-Lord methods, as well as observed and true score equating methods, were employed within the framework of a common-item nonequivalent groups design. To evaluate the equating performance, RMSE and BIAS values were calculated.

Findings: The findings indicate that in a science test with low testlet effects, scale transformation results based on the UIRT model and test equating results based on the bifactor model demonstrated lower error rates. Moreover, as sample size increased, the error in parameter estimations generally decreased, with the TRT model specifically requiring a sample size of at least 500 for robust estimations.

Highlights: The bifactor model, taking testlet effects into account, yielded more precise and consistent results, facilitating fair and reliable score equating. This study, utilizing real data, concretely illustrates the practical implications of testlet effects in tests containing testlets.

Öz

Çalışmanın amacı: Bu çalışmada madde takımları içeren testlerde farklı madde tepki kuramı modelleri ve örneklem büyüklükleri koşullarına dayalı test eşitleme performansları incelenmiştir.

Materyal ve Yöntem: Bu amaçla araştırmada, eTIMSS 2019 bilim testine ait veriler kullanılarak, Tek Boyutlu Madde Tepki Kuramı (TBMTK), Madde Takımı Tepki Kuramı (MTTK) ve bifaktör modelleri altında farklı örneklem büyüklüklerinde yapılan ölçek dönüştürme yöntemleri ve test eşitleme sonuçları incelenmiştir. Denk olmayan gruplarda ortak madde deseni altında ortalama-sigma ve Stocking-Lord ölçek dönüştürme yöntemleri ve gerçek ile gözlenen puana dayalı eşitleme yöntemleri kullanılmıştır. Değerlendirme ölçütleri olarak RMSE ve BIAS değerleri hesaplanmıştır.

Bulgular: Genel olarak düşük düzeyde madde takımı etkisinin olduğu bilim testinde TBMTK modeline dayalı ölçek dönüştürme ve bifaktör modele dayalı test eşitleme sonuçlarının daha düşük hata değerleri ürettiği görülmüştür. Ayrıca örneklem büyüklüğü arttıkça genel olarak parametre kestirimlerinin hata değerlerinin azaldığı gözlemlenmiş olup özellikle MTTK ile çalışıldığında örneklem sayısının 500'den fazla olması gerektiği sonucuna varılmıştır.

Önemli Vurgular: Madde takımı etkisi göz önüne alındığında, bifaktör model daha doğru ve kararlı sonuçlar sunarak adil ve güvenilir puan eşitlemesi yapılmasını sağlamaktadır. Gerçek veri seti kullanılarak gerçekleştirilen bu çalışma ile madde takımları içeren testlerde madde takımı etkisinin pratikte nasıl bir etki yarattığı somut bir şekilde ortaya konulmuştur.

[1] A part of this study was presented at the 9th International Conference on Measurement and Evaluation in Education and Psychology in 2024.

[2] Corresponded Author, Ministry of National Education, İstanbul, Türkiye; https://orcid.org/0000-0001-5671-6858

[3] Hacettepe University, Department of Educational Sciences, Ankara, Türkiye; https://orcid.org/0000-0002-3580-5568

[4] Kütahya Dumlupınar University, Department of Educational Sciences, Kütahya, Türkiye; https://orcid.org/0000-0002-6453-3258

Citation/Alıntı: Dilek, H., Atalay Kabasakal, K., & Gören, S. (2025). Examination of Scale Transformation and Test Equating Methods in Testlet Based Tests, *Kastamonu Education Journal, 33*(3), 658-671. doi: 10.24106/kefdergi.1750267

INTRODUCTION

Assessment in education is a pivotal tool for decision-making in areas such as academic achievement, selection, placement, promotion and access to educational opportunities. In particular, the validity and reliability of large-scale test scores, both at the national and international levels, are of critical importance. For instance, ensuring the security and comparability of items in selection exams administered at different times but serving the same purpose represents a significant challenge. Consequently, multiple parallel test forms are often developed for the same purpose. However, the interchangeability of these forms necessitates empirical evidence demonstrating their equivalence. Test equating encompasses statistical procedures designed to ensure that scores obtained from comparable tests, which are administered for the same purpose and share similar difficulty and content, can be used interchangeably (Kolen & Brennan, 2014). Through test equating, the comparability and validity of scores across different test forms are maintained, thereby supporting fair decision-making processes. However, the equating process can encounter specific challenges in practice. One notable complication arises from the use of testlets, which are commonly employed in large-scale assessments. A testlet refers to a cluster of items that are linked to a shared stimulus, such as a text, graphic, figure or table (Wainer & Kiely, 1987). Testlets offer the advantage of reducing the cognitive load for examinees by enabling multiple items to be developed based on a single stimulus, thereby allowing for the inclusion of more items within a given testing period (Bradlow et al., 199). Nevertheless, this dependence on a common stimulus may violate the assumption of local independence (Lee et al., 2001; Wainer et al., 2007).

The primary issue in equating tests containing testlets stems from the violation of local independence, a foundational assumption in Item Response Theory (IRT). Local independence posits that, when individuals' ability (theta) levels are held constant, their responses to items should be statistically independent of one another (Lord, 1980). In the context of testlets, however, responses to items within the same testlet can influence one another, leading to local dependence. Such dependence can adversely affect the reliability of test scores, as well as the estimation of ability and item parameters (Marais & Andrich, 2008; Wainer & Wang, 2000; Yılmaz Koğar & Kelecioğlu, 2017). Consequently, this issue can undermine the accuracy of statistical analyses, including scale transformation methods and test equating processes, which are inherently reliant on parameter estimation. Thus, it is imperative to examine the testlet effect in assessments containing testlets and to consider its magnitude when determining the most appropriate measurement models. By addressing these challenges, the fairness and validity of score equating can be preserved.

To address the issue of local dependence in testlets, the literature proposes several IRT models beyond traditional approaches that treat testlet items as standalone items. One such approach involves scoring all items within a testlet polytomously to form a single unit, referred to as a "super item," using generalized partial credit model or graded response model (Wainer et al., 2007). However, a significant limitation of this method is the loss of information due to combining multiple items into a single polytomous item. To resolve this issue and explicitly model the dependencies among items within a testlet, the Testlet Response Theory (TRT) model was introduced. TRT was first developed by Bradlow et al. (1999) as an extension of the unidimensional IRT (UIRT) model specifically the two-parameter logistic (2PL) IRT model by incorporating an additional parameter to account for the testlet effect. This enhancement laid the foundation for TRT, allowing for the estimation of a testlet effect parameter (γ) for each testlet, which quantifies the influence of the testlet effect on individual performance (DeMars, 2006). The equation representing the 2PL-TRT model, incorporating the testlet effect parameter into the standard 2PL model, is provided in Equation (1).

$$P(\theta_i, \alpha_i, b_i) = \frac{exp\left(\alpha_i\left(\theta_j - b_i - \gamma_{jd(i)}\right)\right)}{1 + exp\left(\alpha_i\left(\theta_j - b_i - \gamma_{jd(i)}\right)\right)} \qquad (1)$$

When conducting test equating under the TRT framework, multiple ability (θ) and discrimination (a) parameters are estimated based on the general trait measured by the test as well as the specific characteristics of the testlets. Specifically, the number of θ and a parameters estimated corresponds to the number of testlets in addition to the general factor. For instance, in a test comprising three testlets, four distinct θ and a parameters are estimated. However, the test equating process should rely on the general factor parameters, θ_1 and a_1. This is because θ_1 and a_1 parameters are calculated by accounting for the testlet effect, whereas the other parameters are specific to each testlet and only influence responses to items within their respective testlet. This ensures that behaviors specific to each testlet do not generalize across the entire test (Tao & Cao, 2016).

The TRT model can be conceptualized as a constrained form of the bifactor model, which is a multidimensional IRT model positing that each item is associated with both a general (primary) factor and at most one specific (secondary) factor. In the bifactor model originally introduced by Gibbons and Hedeker (1992), there is one general factor and k specific factors. Each item in the test loads onto its corresponding specific factor, while also simultaneously loading onto the general factor (Gibbons & Hedeker, 1992). The model imposes no constraints; items are not influenced by factors other than the general or their designated specific factor. This allows for the independent examination of the effects of the general and specific factors (Rijmen, 2009). For dichotomous items, the bifactor model can be represented as follows (Cai et al., 2011):

$$P(\theta_0, \theta_s) = c_i + (1 - c_i)\frac{\exp(a_{0i}\theta_0 + a_{si}\theta_s + d_i)}{1 + \exp(a_{0i}\theta_0 + a_{si}\theta_s + d_i)} \qquad (2)$$

In this formula, θ_0 represents the general factor, while θ_s represents the specific factor. The parameter c_i denotes the probability of guessing the correct answer, and d_i refers to the intercept parameter of the item. a_{0i} indicates the discrimination parameter associated with the general factor, whereas a_{si} indicates the discrimination parameter associated with the specific factor. Both the bifactor and TRT models address testlet effects in assessments involving testlets, but the selection of the appropriate model depends on the research question, data characteristics, and, most importantly, the magnitude of the testlet effect. In the literature, testlet effects are typically categorized into three levels: low (0.0–0.5), moderate (0.5–1.0) and large (above 1.0) (Cao et al., 2014; Wang et al., 2002).

TRT provides superior performance in tests where violations of local independence are prevalent. For tests with low testlet effects, such violations do not result in significant problems and UIRT may outperform multidimensional IRT models (Chen, 2014; He et al., 2012; Huang et al., 2022; Kim et al., 2019). However, in tests with large testlet effects, bifactor and TRT models are more accurate and are therefore recommended for test equating (Cao et al., 2014; He et al., 2012; Tao & Cao, 2016). However, a critical consideration when applying multidimensional IRT models is their dependence on large sample sizes (Reckase, 2009). Sample size significantly affects the accuracy of parameter estimation during the scaling process, which is a key step in test equating (Hambleton & Cook, 1983; Linacre, 1994; Wang & Liu, 2018). Larger sample sizes enhance the precision of test equating (Cui & Kolen, 2009; Liu & Kolen, 2011; Livingston, 1993; Skaggs, 2005). Consequently, there is a clear need for further research to investigate the role of sample size in studies focusing on multidimensional IRT models for assessments involving testlets.

Since different IRT models address local dependence in distinct ways, this variation may result in different outcomes, even when the same test equating process is employed. Consequently, the primary aim of this study is to examine the equating of scores obtained from tests comprising testlets using different IRT models and sample sizes and to provide a comparative analysis of the results. Therefore, before delving into the analysis phase, an overview of the stages, designs, and methods of test equating is presented.

Stages, Designs, and Methods in Test Equating

When two test forms developed for the same purpose are administered, variations in their means and standard deviations often emerge, rendering direct comparisons between the scores obtained from these forms impossible. To enable comparability, the scores must first be placed on a common scale. Subsequently, equating is applied to allow the scores to be used interchangeably. Test equating refers to the statistical transformation of test scores to account for differences in difficulty between test forms, enabling the scores to be used interchangeably (Kolen & Brennan, 2014). This process, which is critical for student evaluations, requires careful attention at each stage to ensure validity and reliability. The test equating process consists of four main stages: selection of the equating design, administration of the test forms, selection of the equating method, and evaluation of the equating results (Hambleton et al., 1991).

Test equating studies are conducted using various designs, the selection of which depends on the characteristics of the tests to be equated and the testing conditions. Common designs for test equating include the single group design, random groups design, single group design with counterbalancing, and common-item design for non-equivalent groups (CINEG), design (Kolen & Brennan, 2014). Based on the selected design, different test forms are administered to specific groups, followed by the application of a suitable equating method. The choice of equating method is determined by the characteristics of the test forms being equated. Test equating methods are generally developed within two theoretical frameworks: Classical Test Theory (CTT) and Item Response Theory (IRT). CTT-based methods include fundamental approaches such as mean equating, linear equating, and equipercentile equating (Hambleton et al., 1991).

IRT-based equating methods, on the other hand, rely on calibration approaches, which can be categorized into two types: concurrent calibration, where item parameters for all test forms are estimated simultaneously and automatically aligned to the same scale, and separate calibration, where item parameters for each test form are estimated independently (Kolen & Brennan, 2014). Due to practical constraints, separate calibration is often regarded as a safer alternative. In this method, scale transformation is necessary to ensure comparability of the estimated parameters. Scale transformation methods used in IRT-based separate calibration are classified into two categories: moment methods and characteristic curve methods. Moment methods include the mean-mean (MM) and mean-sigma (MS) approaches, while characteristic curve methods include the Stocking-Lord (SL) and Haebara (HB) methods. Detailed explanations of the MS and SL methods used in this study are provided below.

Mean-Sigma Method (MS): In this method, scale transformation coefficients are determined by using the mean and standard deviation of the difficulty parameters for the common items (Marco, 1977). The formulas for these coefficients are provided in Equations (3) and (4).

$$A=\frac{\sigma(b_i)}{\sigma(b_j)} \tag{3}$$

$$B=\mu(b_j)-A.\mu(b_i) \tag{4}$$

$\mu(b_i)$: The mean of the difficulty parameters of the common items in Test *i*.

$\mu(b_j)$: The mean of the difficulty parameters of the common items in Test *j*.

$\sigma(b_i)$: The standard deviation of the difficulty parameters of the common items in Test *i*.

$\sigma(b_j)$: The standard deviation of the difficulty parameters of the common items in Test *j*.

Stocking-Lord Method (SL): This method aims to reduce the differences between the item characteristic curves for the common items. In this method, the total differences between the item characteristic functions, considering both the item difficulty and discrimination parameters, are computed (Stocking & Lord, 1983). The formula is given in Equation (5).

$$\mathrm{SL}_{\mathrm{diff}}(\theta i) =(\textstyle\sum_{j:V} \left[p_{ij}\left(\theta_{Jj};\widehat{a_{Jj}};\widehat{b_{Jj}},\widehat{c_{Jj}}\right) - \sum_{j:V} p_{ij}(\left(\theta_{Jj};\frac{\widehat{a_{IJ}}}{A};A\hat{b}_{Ij} + B\hat{c}_{Ij}\right)\right]^2 \qquad (5)$$

p_{ij}: Item characteristic function for the respondent i and the item j

$\hat{a}_{Jj}$,$\hat{b}_{Jj}$,$\hat{c}_{Jj}$: Item discrimination, difficulty, and guess parameters for the common item j in Scale J

$\hat{a}_{Ij}$,$\hat{b}_{Ij}$, $\hat{c}_{Ij}$: Item discrimination, difficulty, and guess parameters for the common item j in Scale I

j: V: The total is taken over the common items.

The scale transformation methods described are approaches developed to represent the parameters obtained from different test forms on the same scale. When these methods are applied based on UIRT, they may encounter issues related to the local independence assumption in tests containing testlets. To evaluate the performance of individuals who take different test forms, test equating is conducted. The primary distinction between test equating and scale transformation lies in their objectives: test equating provides evidence that the scores obtained from different test forms can be used interchangeably, while scale transformation establishes a relationship between the parameters of the test forms. In other words, while scale transformation connects the parameters of the tests, test equating ensures that individuals' scores can be used interchangeably. Therefore, the methods used in test equating and scale transformation also differ significantly (Ryan & Broocmann, 2009). In this study, two primary approaches to IRT-based equating—true score equating and observed score equating—were employed during the equating processes. True score equating focuses on the potential performance individuals exhibit on different test forms, whereas observed score equating uses the actual test scores obtained by individuals (Kolen & Brennan, 2014). Both methods aim to make comparisons between different test forms possible, but they are grounded in distinct theoretical frameworks and practical approaches. Understanding the differences and applications of these two methods is critical for conducting test equating more accurately and reliably.

True score equating and observed score equating are two fundamental methods used to ensure comparability of scores across different test forms. True score equating allows for the comparison of true scores on different tests independently of individuals' abilities and requires the calculation of all item parameters (Ogasawara, 2003; Kolen & Brennan, 2014). Observed score equating, on the other hand, operates on the distributions of observed correct responses, comparing the number of items answered correctly on each test form (Tao & Cao, 2016; Kolen & Brennan, 2014). Both methods utilize scale transformation techniques such as characteristic curve methods (Stocking-Lord and Haebara) and moment methods (Mean-Mean and Mean-Sigma) (Kolen & Brennan, 2014). As observed in some previous studies, a selective approach was adopted, whereby not all methods were utilized (He et al., 2015; Öztürk Gübeş, 2019; Robitzch, 2024; Yıldırım Seheryeli et al., 2021). In this study, the Stocking-Lord method was preferred for its ability to minimize parametric differences and provide greater flexibility (Asriadi & Retnawati, 2023; Kilmen & Demirtaşlı, 2012; Özdemir & Atar, 2022; Robitzch, 2024; Stocking & Lord, 1983; Uysal & Kilmen, 2016), while the Mean-Sigma method was selected because it takes variances into account, yielding more precise equating results (Kolen & Brennan, 2014; Öztürk Gübeş & Kelecioğlu, 2015).

Purpose of the Study

The necessity of the equating process lies in ensuring that the scores obtained by candidates from exams accurately reflect their true achievement levels. For instance, if a student scores low on a difficult test but high on an easier one, this may lead to misleading decisions about the student. Therefore, test equating processes form the foundation of fair and valid assessments in education. It is crucial that this process is conducted meticulously, particularly in cases where local dependency might arise, such as with testlets.

Researchers have conducted numerous studies addressing local item dependencies in tests composed of testlets and proposed various approaches. However, more studies are needed to investigate the impact of such dependencies on the performance of scale transformation and test equating (Cao et al., 2014; Chen, 2014; He, 2012; Huang et al., 2022; Kim et al., 2019; Tao & Cao, 2016). Evidence on the application of IRT models for scale transformation in tests containing testlets is scarce, warranting further research into their performance. For this reason, different IRT models are considered as a condition in this study. Furthermore, the literature reveals that scale transformation results for common-item designs in non-equivalent groups have not been adequately examined using real data. Since most studies rely on simulation data, the common items typically consist only of testlets. However, in real test applications, common items may include both testlets and standalone items, with some testlet effects being low, others moderate, and some high. Hence, this study, based on a real dataset containing both standalone items and testlets, is anticipated to provide a unique contribution to the literature.

Sample size is also a critical factor influencing the accuracy of parameter estimation and the test equating process (Tao & Cao, 2016). IRT models that account for testlet effects require larger sample sizes due to their inclusion of more parameters (Reckase, 2009). Additionally, Huang et al. (2022) concluded in their study that simpler models such as UIRT demonstrate acceptable error and bias levels with larger sample sizes. Consequently, this study incorporates different sample sizes as a condition when comparing the performance of scale transformation methods for common-item designs in non-equivalent groups using real data from tests containing testlets. In conclusion, the study seeks to answer the following research questions:

1. How do the RMSE and BIAS values of item parameters obtained from different scale transformation methods (MS-SL) change when different IRT models (UIRT-TRT-bifactor) are used under varying sample sizes (500-1000-2000-4279) in the eTIMSS science test (booklets 3 and 4)?

2. How do the RMSE and BIAS values of test equating methods (true score and observed score equating) based on different scale transformation methods (MS-SL) vary in the eTIMSS science test (booklets 3 and 4)?

METHOD

Research Design

This study aims to investigate scale transformation and test equating methods based on different IRT models and sample sizes in tests containing testlets, seeking to obtain detailed insights. In this respect, the study is descriptive in nature (Büyüköztürk et al., 2020).

Study Group

The data for the study were drawn from the student responses to the relevant booklets of the eTIMSS 2019 application across all participating countries. Since sample size is also considered as a condition, random subsamples of 500, 1,000, and 2,000 students were selected from the original dataset of 4,279 students. In line with previous studies suggesting that at least 500 examinees are needed for successful test equating (Spence, 1996) and that larger samples yield more accurate results, analyses were conducted using these sample sizes along with the full sample of 4,279 students (Atalay Kabasakal, 2014; Doğuyurt, 2023; Huang, 2022; Ukşul, 2024).The reason for the selection of these booklets is that when all the booklets with common items were reviewed, these booklets contained the highest number of testlets within the common items.

Data and Equating Design

The study employed a common-item design for non-equivalent groups (CINEG), which is frequently used in scale transformation methods and fits the structure of the data. The equating design is presented in Table 1. The data in this study consist of dichotomous variables scored as 0 or 1.

Table 1. Equating Design

Sample	Booklet 3	Common Items	Booklet 4
group 1	+	+	
group 2		+	+

In the CINEG design presented in Table 1, Booklet 4 is equated to Booklet 3 based on the common items. In this study, Booklet 4 of the science test was equated to Booklet 3. The standalone items, testlets, and total number of items included in the booklets of the science test are shown in Table 2.

Table 2. Number of Items in the Science Test

	Booklet 3	Booklet 4	Common Items
Number of standalone items	25	26	14
Number of two-item testlets	1	2	1
Number of three-item testlets	1	-	-
Number of four-item testlets	1	1	1
Number of seven-item testlets	1	1	1
Total number of items	41	41	27

As shown in Table 2, a total of 27 items are common between the booklets. Of these common items, 14 are standalone items, while the remainder are testlets. Among the testlets, one consists of two items, one consists of four items, and one consists of seven items. The testlet effects for the testlets in the science test are presented in Table 3.

Table 3. Testlet Effects in the Science Test

Testlet	Booklet 3	Booklet 4
T1 (2 items)	0.170	
T2 (2 items)*	0.315	0.220
T3 (2 items)		0.254
T4 (4 items)*	0.535	0.589
T5 (7 items)*	0.397	0.335

T=Testlet ; *=Common Testlet

When the testlet effects of the science test are examined in Table 3, it is observed that in Booklet 3, T1, T2, and T3 have low levels of testlet effects, while T4 demonstrates a moderate level. In Booklet 4, T2, T3, and T5 exhibit low levels of testlet effects, whereas T4 has a moderate testlet effect.

Data Analysis

In the analysis phase, the testlet effects in the science test booklets were first calculated, followed by the calibration of item responses using separate calibration, which yielded parameter estimates. The 2PL models provided a better fit to the data, and since the item parameter estimates in the 3PL model were substantially high, the 2PL model was preferred. Ability (Theta) parameters were estimated using the Expected A Posteriori method, while item parameters were estimated using the Maximum Likelihood Estimation (MLE) method (Embretson & Reise, 2000). The accuracy of the equating results was evaluated using the BIAS value, representing the difference between estimated and true parameter values, and the RMSE value, representing the total equating error. The closer these values are to zero, the more precise the equating results (Kolen & Brennan, 2014). The RMSE formula is provided in Equation (8), while the BIAS formula is provided in Equation (9). In the formula, N represents the total number of individuals, $\bar{x}_I$ denotes the estimated parameter for item i, x_i refers to the true parameter for item i.

$$\text{RMSE}=\sqrt{\frac{\sum_{i=1}^{K}(\bar{x}i-xi)^2}{N}} \tag{6}$$

$$\text{BIAS}=\frac{\sum_{i=1}^{K}(\bar{x}i-xi)}{N} \tag{7}$$

Although RMSE and BIAS values are useful for assessing the overall performance of the equating process, they do not provide information about whether the errors are within acceptable limits. At this stage, the Differences That Matter (DTM) framework (Dorans& Feigenbaum, 1994) offers a more tangible and practical criterion for evaluating the magnitude of errors in practical terms (Kolen & Brennan, 2014; Lee et al., 2012). According to this criterion, if error values are smaller than 0.5, the equating process is considered acceptable. This criterion was used to interpret the error values obtained in this study. Data analysis was conducted using the *mirt* package (Chalmers, 2012) and the *plink* package (Weeks, 2010) in R programming.

FINDINGS

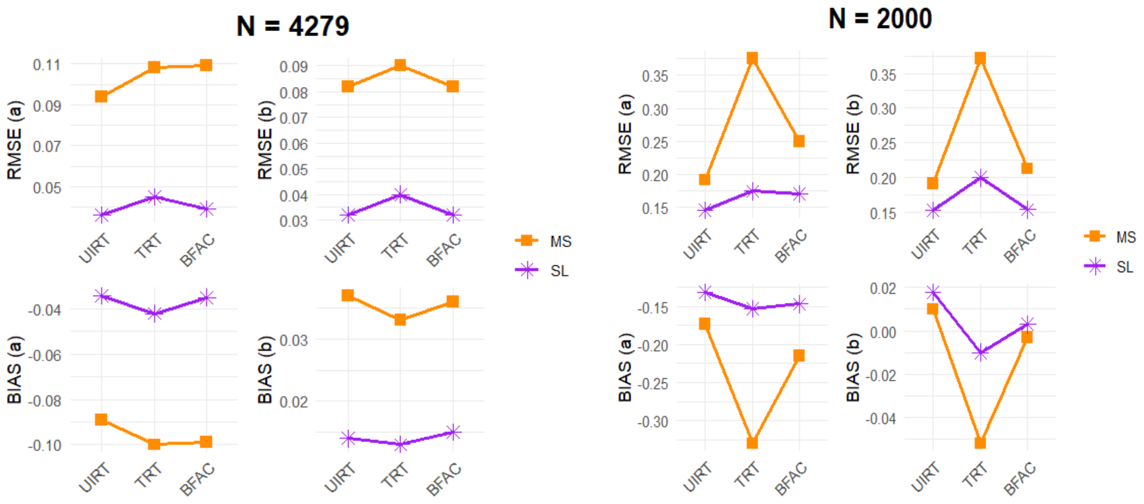
This section presents findings regarding how error and bias values of IRT-based scale transformation methods (MS, SL) vary under different IRT models and sample sizes in the eTIMSS science dataset, which comprises testlets and standalone items from Booklets 3 and 4. Under these conditions, the RMSE and BIAS values related to item parameters are provided in Table 4.

Table 4. Error Values Across All Conditions for Scale Transformation Methods

Scale Transformation Methods	Sample Size	Model	RMSE		BIAS	
			a	b	a	b
Mean-Sigma	4279	UIRT	0.094	0.082	-0.089	0.037
		TRT	0.108	0.09	-0.1	0.033
		Bifactor	0.109	0.082	-0.099	0.036
	2000	UIRT	0.191	0.192	-0.173	0.01
		TRT	0.375	0.371	-0.33	-0.052
		Bifactor	0.25	0.213	-0.215	-0.003

	1000	UIRT	0.215	0.251	0.189	0.016
		TRT	0.249	0.283	0.212	0.033
		Bifactor	0.258	0.273	0.217	0.027
	500	UIRT	0.302	0.265	-0.24	-0.126
		TRT	0.589	0.252	-0.302	-0.123
		Bifactor	0.497	0.304	-0.351	-0.139
Stocking-Lord	4279	UIRT	0.036	0.032	-0.034	0.014
		TRT	0.045	0.04	-0.042	0.013
		Bifactor	0.039	0.032	-0.035	0.015
	2000	UIRT	0.145	0.153	-0.131	0.018
		TRT	0.174	0.2	-0.153	-0.01
		Bifactor	0.17	0.154	-0.147	0.003
	1000	UIRT	0.275	0.338	0.242	0.019
		TRT	0.327	0.394	0.278	0.035
		Bifactor	0.332	0.373	0.28	0.039
	500	UIRT	0.189	0.227	-0.15	-0.163
		TRT	0.4	0.238	-0.205	-0.177
		Bifactor	0.288	0.248	-0.204	-0.179

When the error values of item parameter estimate for each scale transformation method presented in Table 4 are examined under the conditions of IRT models and sample sizes, it is observed that the UIRT, bifactor, and TRT models consistently yield the lowest error values, respectively. Moreover, a general trend indicates that as the sample size decreases, the error values of parameter estimates increase. While the evaluation criteria employed during the equating process (e.g., BIAS, RMSE) are particularly practical for comparative analyses, they remain insufficient in determining the acceptability of the results. Accordingly, when assessed using the DTM criterion, it is evident that, among the MS methods, all values—except for the RMSE value of the TRT discrimination parameter with a sample size of 500—fall below 0.5, signifying their acceptability. To facilitate interpretation and enable a more detailed analysis, a line graph illustrating the parameter error values estimated under varying sample size and IRT model conditions is presented in Figure 1.



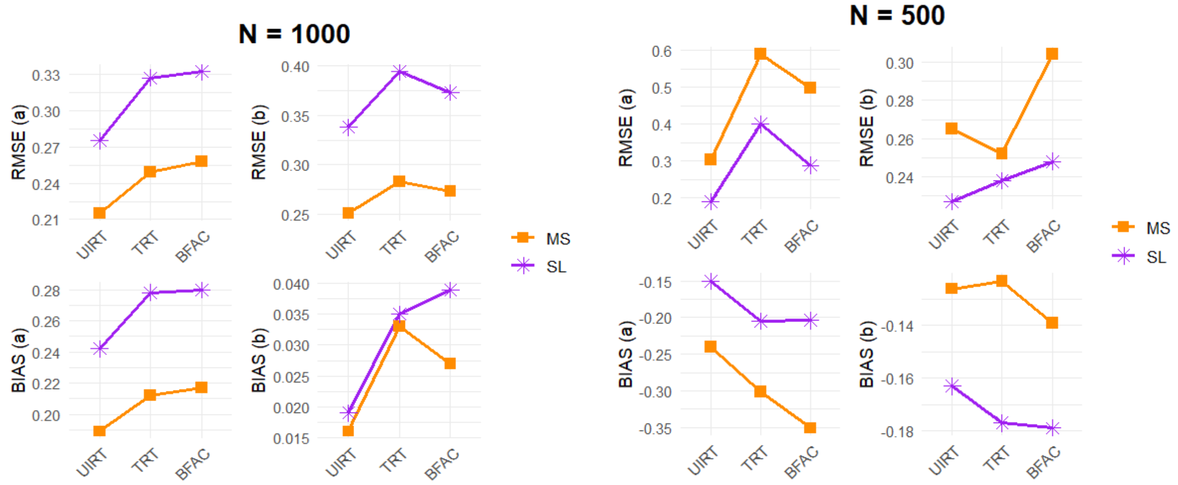


Figure 1. Line Graph of Error Values Under All Conditions According to Scale Transformation Methods

A review of Figure 1 reveals that parameter estimation error values are predominantly the highest for the MS method (except for N=1000). In contrast, the results derived from the SL methods are observed to be significantly similar across most conditions and consistently lower than those obtained from the MS method. Notably, all computed values remain below the threshold of 0.5, signifying their acceptability. Moreover, a model-specific analysis indicates that the UIRT model yields the lowest error values for both the a and b parameters. In cases involving larger sample sizes, the bifactor model exhibits error values comparable to those of the UIRT model. Additionally, the BIAS values across all conditions are found to be exceptionally low.

For the real dataset featuring the optimal sample size (n=4279), descriptive statistics of the common item parameters, as well as the equating constants corresponding to the scale transformation methods, have been computed for both test forms and are summarized in Table 5.

Tablo 5. Descriptive Statistics of Common Item Parameters and Equating Constants by Forms

	IRT				TRT				Bifactor			
	Form X		Form Y		Form X		Form Y		Form X		Form Y	
	a	b	a	b	a	b	a	b	a	b	a	b
Mean	1.123	-0.629	1.098	-0.605	1.162	-0.664	1.135	-0.647	1.229	-0.741	1.251	-0.79
Sd	0.351	0.989	0.357	1.072	0.393	1.038	0.415	1.134	0.485	1.165	0.550	1.57
		MS	SL			MS	SL			MS	SL	
A		0.922	0.969			0.915	0.962			0.742	0.937	
B		-0.070	-0.027			-0.071	-0.032			-0.154	-0.068	

When Table 5 is examined, it is observed that the mean values of the a and b parameters for the common items are consistent across all models, with the equating constants being nearly identical in both the IRT and TRT models.

Descriptive statistics for the parameter values of the common items were calculated for each model using two scale transformation methods MS (moment methods) and SL (characteristic curve methods) in the scaled form.

Table 6. Descriptive Statistics of Item Parameters in Scaled Form

	UIRT				TRT				Bifactor			
	MS		SL		MS		SL		MS		SL	
	a	b	a	b	a	b	a	b	a	b	a	b
Mean	1.100	-0.490	1.047	-0.467	1.135	-0.647	1.177	-0.653	1.251	-0.79	1.229	-0.741
Sd	0,402	0.783	0.383	0.823	0.415	1.134	0.431	1.091	0.550	1.57	0.485	1.165

In Table 6, it is also apparent that the mean values of the common item parameters are quite similar in both the UIRT and TRT models.

A test equating procedure was performed based on both observed and true scores using the SL (characteristic curve methods) and MS (moment methods) scale transformation methods on the real dataset. Weighting was applied during observed score equating. Specifically, ten ability (theta) points were selected within the range of [-4, 4], as recommended by Kolen and Brennan (2014). Subsequently, rescaled theta points were obtained based on the A and B constants derived from the scale transformation methods. Additionally, the standard normal distribution density was computed.

The observed score equating was performed using the rescaled theta points and weights, with the results presented in Table 7.

Table 7. RMSE and BIAS Values Based on Observed and True Score Equating

Equating Method	Model	Mean-Sigma		Stocking Lord	
		RMSE	BIAS	RMSE	BIAS
Observed Score-Based	UIRT	0.795	-0.326	0.467	-0.199
	TRT	0.812	-0.313	0.495	-0.277
	Bifactor	0.769	-0.299	0.413	-0.227
True Score-Based	UIRT	0.808	-0.260	0.471	-0.153
	TRT	0.824	-0.249	0.493	-0.161
	Bifactor	0.786	-0.237	0.439	-0.152

Upon reviewing Table 7, it is evident that the best results were obtained for the bifactor model in both observed and true score equating methods. The RMSE values for all models, derived from the MS method, were greater than 0.5 in both equating methods, thus failing to meet the DTM criterion. On the other hand, the BIAS values remained below 0.5 across all scale transformation and equating methods.

Figure 2 presents line graphs depicting the variation of error values for observed and true scores based on the MS and SL scale transformation methods. In the UIRT model, the RMSE values resulting from both observed and true score equating were nearly identical for both the MS and SL scale transformation methods. In the TRT model, although the RMSE values were similar, the bifactor model displayed differences, with the MS method producing higher RMSE values. When examining the RMSE and BIAS values, it can be concluded that the SL scale transformation method produced lower error values across all models.

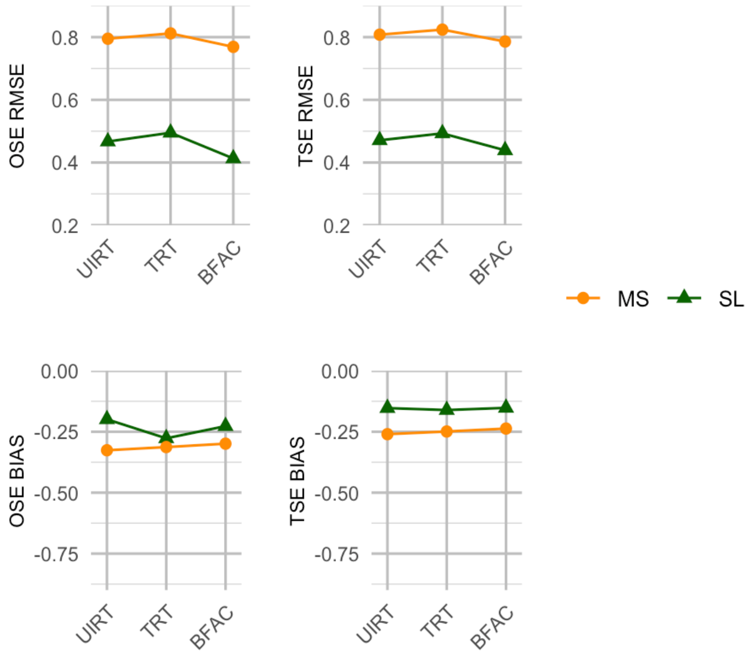


Figure 2. Line Graph of Error Values Based on OSE and TSE Equating

Table 8 presents the averages and standard deviations of the original forms and the scores equated from Form 4 to Form 3, based on the observed and true scores for all three models.

Table 8. Moments for equating Form X and Y

	Form	UIRT		TRT		Bifactor	
		Mean	Sd	Mean	Sd	Mean	Sd
True Score	Equating	20.653	12.579	20.661	12.586	20.652	12.575
Observed Score	Equating	20.699	12.595	20.778	12.602	20.727	12.566
Original Score	Booklet 3	23.925	7.576				
	Booklet 4	24.190	7.701				

As seen in Table 8, using the IRT true score equating, the average score for Form 4 equated to the 3rd Form is 20.653; when using the IRT observed score equating, the average is 20.699. For TRT, these scores are 20.661 and 20.778, respectively. The values obtained for the bifactor model are quite similar to those for the UIRT model, with values of 20.652 and 20.727, respectively. The moments of the transformed scores are very similar for the two IRT methods. Furthermore, without using the IRT model, the averages for the original Forms 3 and 4 are 23.925 and 24.190, respectively.

RESULTS AND DISCUSSION

This study compares the error values (RMSE-BIAS) obtained from different scale transformation methods (MS-SL) and equating methods (OSE-TSE) under various conditions, including different sample sizes (500-1000-2000-4279) and IRT models (UIRT-TRT-bifactor), for equating the 3rd and 4th Forms of the eTIMSS 2019 science test based on a common test design across unequal groups. The results of the study show that when the testlet effect in the science test is at a medium or low level, the scale transformations based on UIRT yield lower RMSE and BIAS values for parameter errors. This finding is consistent with previous studies in the literature, which have shown that the UIRT model provides better results in datasets with a low testlet effect (Chen, 2014; He, 2012; Huang et al., 2022; Kim et al., 2019; Tao & Cao, 2016). When comparing the bifactor and TRT models, it was found that the error values obtained from the bifactor model were lower. In fact, in large samples, the error values obtained from bifactor and UIRT models were quite similar. This suggests that bifactor models better model the local independence violations caused by testlet effects and should be preferred for scale transformation studies in tests where the testlet effect is present.

When examining the research findings related to sample size, under the CİNEG design, errors in parameter estimation for scale transformation methods generally decreased as sample size increased in all models. This result aligns with the information in the literature, which indicates that equating procedures require a large sample size for accurate parameter estimation (Babcock & Hodge, 2019; Huang et al., 2022; Kolen & Brennan, 2014). Specifically, for datasets with a sample size of 500, the RMSE value can be high when using the TRT model. This is likely due to the TRT model treating each testlet as a separate dimension, which necessitates larger sample sizes in analyses (Wainer et al., 2007). Therefore, when using the 2PL-TRT model for tests composed entirely of testlets, the sample size should be greater than 500.

In all models and sample sizes, the highest parameter estimation error values were generally observed with the MS method. The SL method, on the other hand, produced lower errors compared to the MS method. This finding is consistent with the results of Gök and Kelecioğlu (2014) and Ogasawara (2001) but contradicts the findings of Gül et al. (2017) and Kim & Lee (2006). In true score equating and observed score equating, among the two scale transformation methods used (MS-SL), the MS method again resulted in higher RMSE values across all models. The reason for this outcome in true and observed score equating may be attributed to the more erroneous parameter estimations made by the MS method compared to the SL method during the scale transformation process. Research comparing characteristic curve methods (SL) with moment methods (MS) for binary IRT models has demonstrated that characteristic curve methods yield more stable results (Baker & Al-Karni, 1991; Gül et al., 2017; Hanson & Béguin, 2002; Kim & Cohen, 1991; Kolen & Brennan, 2014; Lee & Ban, 2010; Li et al., 2012; Ogasawara, 2001; Zor, 2023).

True and observed score equating were conducted with a sample size of 4279 using the SL and MS scale transformation methods. Most of the error values in this case were found to be below the DTM criterion, with the results of equating using only the MS scale transformation method showing errors exceeding 0.5. Furthermore, the bifactor model provided the best results in both true and observed score equating. Therefore, the use of the bifactor model based on the SL scale transformation method in equating studies may be appropriate. The literature also suggests that when the testlet effect is low, the bifactor model is a suitable option for test equating (He et al., 2012; Kim et al., 2019). Additionally, the results obtained align with the findings of He & Li (2014), where the bifactor model produced the lowest error in equating studies. Thus, in tests predominantly or entirely composed of testlets, where the testlet effect needs to be considered, using the bifactor model with an adequate sample size would be more appropriate.

Upon examining the results of observed and true score equating, it was found that the equating based on observed scores generally produced lower error values. Only the equating results for the TRT model using the SL scale transformation method were found to be quite similar. However, studies in the literature have shown that true and observed score equating based on the UIRT model yield indistinguishable results (Cao et al., 2014; Han et al., 1997; Lord & Wingersky, 1984).

A significant contribution of this study is that, although most studies in the literature regarding testlets have been conducted using simulation data (Cao et al., 2014; Chen, 2014; Huang et al., 2022; Kim et al., 2019; Tao & Cao, 2016), this study was carried out with real data. Unlike simulation data, the use of real data offers a more tangible understanding of how the testlet effect and violations of local item independence affect the results in practice. These results underscore the importance of controlling for the testlet effect in testlet based tests and highlight the applicability of the UIRT model in cases where the testlet effect is low or when most of the items in the test are standalone items.

In the common-item nonequivalent groups (CINEG) design, the proportion of common items (i.e., the ratio of common items to the total number of items) is a critical factor in test equating. As a general guideline, the literature recommends that the proportion of common items should be at least 20%. This proportion is recognized as one of the key factors affecting the accuracy of equating, and it is particularly advised to maintain a higher proportion when the total number of items in the test is relatively small (Kolen & Brennan, 2014). In the present study, the proportion of common items exceeds the recommended range, which is considered advantageous for enhancing the accuracy of the equating process. However, this research has some important limitations. Since the study was conducted using a real dataset, the number of testlets in the eTIMSS science test for Forms 3 and 4 was considerably lower than the number of standalone items, and the testlet effect was low. Therefore, future studies should focus on tests composed entirely of testlets or with a similar number of standalone and testlet items. Additionally, working with real datasets exhibiting a strong testlet effect will provide clearer insights into the impact of local dependence on the models. Furthermore, research addressing conditions such as sample size, the number of testlet items, standalone items, testlet effects and the ability distribution (kurtosis-skewness) of the dataset, whether based on real or simulation data, would provide significant contributions to the literature.

Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The authors received no financial support for the research, author-ship, and/or publication of this article.

Statements of publication ethics

We hereby declare that the study has not unethical issues and that research and publication ethics have been observed carefully.

Researchers’ contribution rate

The study was conducted and reported with equal collaboration of the researchers.

Ethics Committee Approval Information

In this study, the eTIMSS 2019 dataset, which is directly downloadable from the TIMSS 2019 International Database website, was used. Since no data collection process was involved in obtaining open-access dataset, ethical committee approval was not required.

References

Asriadi M., H. (2023). Equating of standardized science subjects tests using various methods: which is the most profitable? *Thabiea : Journal of Natural Science Teaching, 6*(1), 51-64.

Atalay Kabasakal, K. (2014). *Değişen madde fonksiyonunun test eşitlemeye etkisi* [Doktora tezi]. Hacettepe Üniversitesi.

Babcock, B., & Hodge, K. J. (2020). Rasch versus classical equating in the context of small sample sizes. *Educational and Psychological Measurement*, *80*(3), 499-521. https://doi.org/10.1177/0013164419878

Baker, F. B. & Al-Karni, A. (1991). A comparison of two procedures for computing IRT equating coefficients. *Journal of Educational Measurement, 28*(2), 147-162 https://doi.org/10.1111/j.1745-3984.1991.tb00350.x

Bradlow, E. T., Wainer, H., & Wang, X. (1999). A Bayesian random effects model for testlets. *Psychometrika, 64*(2), 153–168. https://doi.org/10.1007/BF02294533

Büyüköztürk, Ş., Kılıç Çakmak, E., Akgün, Ö. E., Karadeniz, Ş., & Demirel, F. (2020). *Eğitimde bilimsel araştırma yöntemleri*. Pegem Akademi.

Cai, L., Yang, J. S., & Hansen, M. (2011). *Generalized full-information item bifactor analysis*. Psychol Methods, 16(3), 221–248. 10.1037/a0023350

Cao, Y., Lu, R., & Tao, W. (2014). *Effect of item response theory (IRT) model selection on testlet- based test equating* (ETS Research Report No. RR-14-19). Educational Testing Service. https://doi.org/10.1002/ets2.12017

Chalmers, R. P. (2012). mirt: A Multidimensional Item Response Theory Package for the R Environment. *Journal of Statistical Software, 48*(6), 1–29. https://doi.org/10.18637/jss.v048.i06

Chen, J. (2014). *Model selection for IRT equating of testlet-based tests in the random groups design* [Doctoral dissertation] The University of Iowa. ProQuest Dissertations Publishing.

Cui, Z., & Kolen, M. J. (2009). Evaluation of two new smoothing methods in equating: The cubic B-spline presmoothing method and the direct presmoothing method. *Journal of Educational Measurement*, *46*(2), 135-158. https://doi.org/10.1111/j.1745-3984.2009.00074.x

DeMars, C. (2006). Application of the bi-factor multidimensional item response theory model to testlet-based tests. *Journal of Educational Measurement, 43*(2), 145–168. https://doi.org/10.1111/j.1745-3984.2006.00010.x

Doğuyurt, A. (2023). *İkili puanlanan testlerde yerel madde bağımsızlık varsayımının ihlâlinin test eşitleme yöntemlerine etkisi* [Doctoral dissertation]. Gazi Üniversitesi.

Dorans, N. J., & Feigenbaum, M. D. (1994). *Equating issues engendered by changes to the SATR and PSAT/NMSQTR* (ETS RM-94-10). ETS.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Lawrence Erlbaum Associates.

Gibbons, R. D., & Hedeker, D. (1992). Full information item bifactor analysis. *Psychometrika*, 57(3), 423-436. https://doi.org/10.1007/BF02295430

Gök, B., & Kelecioğlu, H. (2014). Denk olmayan gruplarda ortak madde deseni kullanılarak madde tepki kuramına dayalı eşitleme yöntemlerinin karşılaştırılması. *Mersin Üniversitesi Eğitim Fakültesi Dergisi, 10*(1), 120-136. https://doi.org/10.17860/efd.78698

Gül, E., Doğan-Gül, Ç., Çokluk-Bökeoğlu, Ö. & Özkan, M. (2017). Temel eğitimden ortaöğretime geçiş matematik alt testi asıl sınav ve mazeret sınavlarının madde tepki kuramına göre eşitlenmesi. *Abant İzzet Baysal Üniversitesi Eğitim Fakültesi Dergisi, 17*(4), 1900-1915. https://doi.org/10.17240/aibuefd.2017.17.32772-363973

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

Hambleton, R.K., & Cook, L.L. (1983). Robustness of ítem response models and effects of test length and sample size on the precision of ability estimates. In D.J. Weiss (Ed.), *New horizons in testing: Latent trait test theory and computerized adaptive testing* (pp. 31-49). Vancouver.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Kluwer.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. SAGE Publications.

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

Hanson, B. A. & Béguin, A. A. (2002). Obtaining a common scale for Item Response Theory item parameters using separate versus concurrent estimation in the common-item equating design. *Applied Psychological Measurement, 26*(1), 3-24. https://doi.org/10.1177/0146621602026001001

He, W., Li, F., Wolfe, E. W., & Mao, X. (2012). Model selection for equating testlet-based tests in the NEAT design: An empirical study. *Annual NCME Conference*.

He, Y., Zhongmin, C. & Osterlind, S. J. New robust scale transformation methods in the presence of outlying common items. *Applied Psychological Measurement 39* (8), 613-626. https://doi.org/10.1177/0146621615587003

Huang, F., Li, Z., Liu, Y., Su, J., Yin, L., & Zhang, M. (2022). An extension of testlet-based equating to the polytomous testlet response theory model. *Frontiers in Psychology, 12*, 743362. https://doi.org/10.3389/fpsyg.2021.743362

Kilmen, S. &, Demirtaşlı, N. (2012). Comparison of test equating methods based on item response theory according to the sample size and ability distribution. *Procedia - Social and Behavioral Sciences 46*, 130 – 134. 10.1016/j.sbspro.2012.05.081

Kim, S. & Cohen, A. S. (1991). Effects of linking methods on detection of DIF. *Journal of Educational Measurement, 29*(1), 51-66.

Kim, S. & Lee, W. (2006). An extension of four IRT linking methods for mixed-format tests. Journal of Educational *Measurement, 43*(1), 53-76.

Kim, K. Y., Lim, E., & Lee, W. C. (2019). A comparison of the relative performance of four IRT models on equating passage-based tests. *International Journal of Testing, 19*(3), 248–269. https://doi.org/10.1080/15305058.2018.1530239

Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scaling, and linking: Methods and practices* (3rd ed.). Springer.

Lee, W. C., and Ban, J. C. (2009). Comparison of IRT linking procedures. *Applied Measurement in Education, 23*(1), 23-48. https://doi.org/10.1080/08957340903423537

Lee, W., He, Y., Hagge, S., Wang, W., & Kolen, M. J. (2012). Equating mixed-format tests using dichotomous common items. In M. J. Kolen & W. Lee (Eds.), *Mixed-format tests: Psychometric properties with a primary focus on equating* (Vol. 2). CASMA Monograph. Iowa City, IA: Center for Advanced Studies in Measurement and Assessment, The University of Iowa.

Lee, G., Kolen, M. J., Frisbie, D. A., & Ankenmann, R. D. (2001). Comparison of dichotomous and polytomous item response models in equating scores from tests composed of testlets. *Applied Psychological Measurement, 25*, 357–372. https://doi.org/10.1177/01466210122032226

Linacre, J. M. (1994). Sample size and item calibration stability. *Rasch measurement transactions*, *7*, 328.

Liu, C., & Kolen, M. J. (2011). A comparison among IRT equating methods and traditional equating methods for mixed-format tests. *Mixed-format tests: Psychometric properties with a primary focus on equating*, *1*, 75-94.

Livingston, S. A. (1993). Small-sample equating with log-linear smoothing. *Journal of Educational Measurement, 30*(1), 23–29.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Lawrence Erlbaum Associates.

Lord, F. M. & Wingersky, M. S. (1984). Comparison of IRT true score and equipercentile observed score “equatings.”. *Applied Psychological Measurement*, *8*, 453–46.

Loyd, B. H., & Hoover, H. D. (1980). Vertical equating using the Rasch model. *Journal of Educational Measurement, 17*(3), 179–193.

Marais, I. D., & Andrich, D. (2008). Effects of varying magnitude and patterns of local dependence in the unidimensional Rasch model. *Journal of Applied Measurement, 9*(2), 105–124.

Marco, G. L. (1977). Item characteristic curve solutions to three intractable testing problems. *Journal of Educational Measurement, 14*(2), 139–160.

Ogasawara, H. (2000). Asymptotic standard errors of IRT equating coefficients using moments. *Economic Review, Otaru University of Commerce, 51*(1), 1–23.

Ogasawara, H. (2001). Least squares estimation of item response theory linking coefficients. *Applied Psychological Measurement*, *25*, 3–24.

Ogasawara, H. (2003). Asymptotic standard errors of IRT observed-score equating methods. *Psychometrika, 68*(2), 193–211.

Özdemir, G., & Atar, B. (2022). Investigation of the missing data imputation methods on characteristic curve transformation methods used in test equating. Journal of Measurement and Evaluation in Education and Psychology, 13(2), 105-116. https://doi.org/10.21031/epod.1029044Öztürk Gübeş, N. &, Kelecioğlu, H. (2016). The impact of test dimensionality, common-item set format, and scale linking methods on mixed format test equating. *Educational Sciences: Theory & Practice, 16*(3), 715-734. 10.12738/estp.2016.3.0218

Öztürk Gübeş, N. (2019). Test eşitlemede çok boyutluluğun eş zamanlı ve ayrı kalibrasyona etkisi. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi, 34*(4), doi: 1061-1074. 10.16986/HUJE.2019049186

Reckase, M. D. (2009). *Multidimensional item response theory*. Springer.

Rosenbaum, P. R. (1988). Items bundles. *Psychometrika, 53*(3), 349–359.

Rijmen, F. (2009). Three multidimensional models for testlet-based tests: Formal relations and an empirical comparison. *ETS Research Report, 2009*(2), 1–41.

Robitzsch, 2024. Bias-reduced Haebara and Stocking–Lord linking. *J Multidisciplinary Scientific journal, 7*(3), 373-384. https://doi.org/10.3390/j7030021

Ryan, J., & Brockmann, F. (2009). *A Practitioner’s Introduction to Equating with Primers on Classical Test Theory and Item Response Theory*. Council of Chief State School Officers.

Skaggs, G. (2005). Accuracy of random groups equating with very small samples. *Journal of Educational Measurement, 42*(4), 309-330. https://doi.org/10.1111/j.1745-3984.2005.00018.x

Spence, P. (1996). *The effect of multidimensionality on unidimensional equating with item response theory* [Doctoratal dissertation] University of Florida.

Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement, 7*(2), 201–210.

Tao, W., & Cao, Y. (2016). An extension of IRT-based equating to the dichotomous testlet response theory model. *Applied Measurement in Education, 29*(2), 108–121. https://doi.org/10.1080/08957347.2016.1138956

Uşkul, B. (2024). *Madde takımı tabanlı testlerde ölçek dönüştürme hatalarının incelenmesi* [Doctoral dissertation]. Hacettepe Üniversitesi.

Uysal, İ. &, Kilmen, S. (2016). Comparison of item response theory test equating methods for mixed format tests. *International Online Journal of Educational Sciences, 2016, 8* (2), 1-11. http://dx.doi.org/10.15345/iojes.2016.02.001

Wainer, H., & Kiely, G. L. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational Measurement, 24*(3), 185–202.

Wainer, H., & Wang, X. (2000). Using a new statistical model for testlets to score TOEFL. *Journal of Educational Measurement, 37*(3), 203–220. https://doi.org/10.1111/j.1745-3984.2000.tb01083.x

Wainer, H., Bradlow, E. T., & Wang, X. (2007). *Testlet response theory and its applications*. Cambridge University Press.

Wang, X., Bradlow, E. T., & Wainer, H. (2002). A general Bayesian model for testlets: Theory and applications. *Applied Psychological Measurement, 26*(2), 190–218. https://doi.org/10.1177/0146621602026001007

Wang, S. & Liu, H. (2018). Minimum sample size needed for equipercentile equating under the random groups design. In M. J. Kolen ve W. Lee (Ed.), *Mixed-format tests: Psychometric properties with a primary focus on equating* (vol 2.5, s. 107-126). Center for Advanced Studies in Measurement and Assessment.

Weeks, J. P. (2010). plink: An R package for linking mixed-format tests using IRT-based methods. *Journal of Statistical Software, 35*(12), 1–33.

Yıldırım Seheryeli, M., Yahşi-Sarı, H., & Kelecioğlu, H. (2021). Comparison of Kernel Equating and Kernel Local Equating in Item Response Theory Observed Score Equating. *Journal of Measurement and Evaluation in Education and Psychology*, *12*(4), 348-357. https://doi.org/10.21031/epod.900843

Yılmaz Koğar, E., & Kelecioğlu, H. (2017). Examination of different item response theory models on tests composed of testlets. *Journal of Education and Learning, 6*(4), 113-126. 10.5539/jel.v6n4p113

Zor, Y. M. (2023). Investigation of multidimensional scale transformation methods applied to multidimensional tests according to various conditions. *Adıyaman University Journal of Educational Sciences, 13*(1), 41-53. http://dx.doi.org/10.17984/adyuebd.1239198