International Journal of Assessment Tools in Education
2025, Vol. 12, No. 3, 629–661
https://doi.org/1.21449/ijate.1562627
journal homepage: https://dergipark.org.tr/en/pub/ijate
Research Article

Examining the impact of violations of local item independence assumption on test equating methods

Mehmet Fatih Doğuyurt[1*], **Şeref Tan**[2]

[1]Tokat Gaziosmanpaşa University, Faculty of Education, Department of Educational Sciences, Tokat, Türkiye
[2]Gazi University, Faculty of Education, Department of Educational Sciences, Retired, Ankara, Türkiye

ARTICLE HISTORY
Received: Oct. 7, 2024
Accepted: Feb. 21, 2025

Keywords:
Test equating,
Multidimensional IRT,
Local item dependence,
Equating error.

Abstract: This study investigates the impact of violating the local item independence assumption by loading certain items onto a second dimension on test equating errors in unidimensional and dichotomous tests. The research was designed as a simulation study, using data generated based on the PISA 2018 mathematics exam. Analyses were conducted under 36 different conditions, varying by sample sizes (250, 1000, and 5000), test lengths (20, 40, and 60 items), and proportions of items loaded onto the second dimension (0%, 15%, 30%, and 50%). A "random groups design" was used, resulting in the creation of 3600 datasets through 100 replications. The results revealed that the equating methods based on classical test theory (CTT) showed varying levels of error depending on the error types and conditions. Among the item response theory (IRT) scale transformation methods, the Stocking-Lord method produced the least error values and was the least affected by violations of the local independence assumption. Additionally, the observed score equating method demonstrated lower root mean square error (RMSE) values than the true score equating method and was less affected by local independence violations. The SS-MIRT observed score equating method yielded lower RMSE values compared to the other methods and was found to be more robust against the violation of the local independence assumption.

1. INTRODUCTION

In testing, to prevent students from engaging in cheating or from recalling and answering questions based on previous tests, different test forms with identical characteristics are developed to better assess students' actual performance. These tests, referred to as parallel or alternative forms, aim to measure the same latent trait or construct and are characterized by having the same true score and error variance (De Gruijter & Leo, 2007). Although these tests are assumed to be parallel, it is emphasized that developing truly parallel forms is extremely challenging in practice (Aiken, 2020; Hambleton *et al.*, 1991).

Parallel tests may exhibit minor differences, and the method used to account for these discrepancies is known as test equating (Kolen & Hendrickson, 2013). Test equating is a statistical process that adjusts the differences between parallel tests, ensuring that scores obtained from these tests can be used interchangeably (Kolen & Brennan, 2014). It also involves

*CONTACT: Mehmet Fatih DOĞUYURT ✉ doguyurtmfatih@gmail.com Tokat Gaziosmanpaşa University, Faculty of Education, Department of Educational Sciences, Tokat, Türkiye

e-ISSN: 2148-7456

converting the unit systems of parallel tests to one another, a necessary step for the tests to be deemed equivalent (Angoff, 1984). Within the framework of Classical Test Theory (CTT), test equating methods are categorized into three types: mean equating, linear equating, and equipercentile equating (Kolen & Brennan, 2014). Mean equating is the least restrictive (Sansivieri *et al.*, 2017) and the simplest method (Finch *et al.*, 2014), which focuses on the averages of the tests being equated. This equating method operates on the assumption that if the means of the tests are equal, the scores obtained from these tests will also be equal, with score differences attributed to variations in the difficulty levels of the tests.

The linear equating method (Crocker & Algina, 1986), appropriate when the score distributions of the X and Y forms differ only in terms of mean and standard deviation, is a method that adjusts for differences in test difficulty based on the score scale (Kolen & Brennan, 2014). In linear equating, scores that are the same number of standard deviations away from the means of the tests are equated to be equal (Kolen & Brennan, 2014). The fundamental assumption underlying linear equating is that the score distributions are similar, except for the differences in the means and standard deviations (Crocker & Algina, 1986). It has been argued that applying this method in exams with participants of similar ability levels yields more accurate equating results (Donlon, 1984).

Equipercentile equating seeks to identify which scores on the tests being equated have the same percentile rank (Crocker & Algina, 1986). If a score on the new form and a score on the reference form hold the same percentile rank within the group, these scores are considered equivalent for that group of test takers (Livingston, 2014). Since the individuals to whom the test forms being equated are administered represent a sample drawn from a particular population, the raw score distributions may appear irregular when graphed due to random error (Cui, 2006). Additionally, the absence of scores corresponding to each percentile in the distribution may contribute to irregularities in the score distribution (Kolen & Brennan, 2014). For these reasons, random error is present when estimating equating relationships between test forms' scores. One method to minimize random error is using smoothing techniques. Smoothing is a process that adjusts sample distributions to resemble the population distribution more closely (Kolen & Brennan, 2014; Lim, 2016). Smoothing can be performed in two ways: pre-smoothing and post-smoothing methods. Another set of commonly used equating methods in test equating are Item Response Theory (IRT)-based equating methods. The first step in IRT-based equating is to estimate the item and ability parameters of the test forms. Since the parameters obtained from different forms must be placed on the same scale, a scale transformation is required first (Kolen & Brennan, 2014).

The underlying principle of scale transformation is to align the item and ability parameter estimates obtained from both Form X and Form Y on the same scale. This is accomplished by converting the item and ability parameters estimated from Form X data to the scale of the parameters estimated from Form Y data. To achieve this, it is first necessary to calculate the slope and intercept constants, known as linking coefficients. The slope constant is denoted by A, and the intercept constant is denoted by B. When the separate calibration method is applied to estimate the item and ability parameters of the test forms to be equated, the corresponding value of the ability level (θ) for person i in Form I on Form J is obtained as follows (Kolen & Brennan, 2014):

$$Q_{Ji} = AQ_{Ii}+B \quad (1)$$

Q_{Ji} =The ability level of person i in Form J

Q_{Ii} = The ability level of person i in Form I

A: Slope constant

B: Intercept constant

In addition to the transformation of ability levels, item parameters (item difficulty and item discrimination) are transformed using the Equations 2 and 3 provided below.

$$a_{Ji} = \frac{a_{Ij}}{A} \quad (2)$$

$$b_{Jj} = Ab_{Ij} + B \quad (3)$$

a_{Ji} = The item discrimination parameter of item *j* in Form *J*

a_{Ij} = The item discrimination parameter of item *j* in Form *I*

b_{Jj} = The difficulty parameter of item *j* in Form *J*

b_{Ij} = The difficulty parameter of item *j* in Form *I*

Since the lower asymptote parameter, i.e., the *c* parameter, is on a probability scale, no transformation is applied (Kolen & Brennan, 2014). That is, the *c* parameter remains constant for both forms and is symbolized as shown in Equation 4.

$$c_{Jj}=c_{Ij} \quad (4)$$

The most commonly used scale transformation methods in the literature are the mean/sigma method (Marco, 1977), mean/mean method (Loyd & Hoover, 1980), Haebara method (Haebara, 1980), and Stocking-Lord method (Stocking & Lord, 1983) (Kolen & Brennan, 2014). The first two methods are referred to as "moment methods", while the remaining ones are called "characteristic curve methods".

The mean-sigma method, defined by Marco (1977), is also referred to as the mean-standard deviation method. In the mean-sigma method, the average and standard deviation of the difficulty parameter are used to obtain the slope (*A*) and intercept (*B*) constants. In the mean-mean transformation method, defined by Loyd and Hoover (1980), the average of the item difficulty and item discrimination parameters of the test forms to be equated is used (Kolen & Brennan, 2014). Baker and Al-Karni (1991) and Ogasawara (2000) stated that the mean-mean method provides more stable results than the mean-sigma method, as the means are more stable than the standard deviations, and thus, it can be preferred over the mean-standard deviation method. Kolen and Brennan (2014) suggested that the mean-sigma method might sometimes be preferred over the mean-mean method due to the more stable estimates of the *b* parameter compared to the *a* parameter estimates. They recommended that equating should be done using both methods and that the raw/scale score transformations obtained from both methods should be compared.

An issue related to moment methods is that the item parameter estimates may generate nearly identical item characteristic curves (ICCs) across the ability range created by the test takers' scores. However, for an item with different item difficulty parameter values and similar ICCs, the mean-sigma method will be influenced by differences in the *b* parameter estimates. The primary cause of this issue is that moment methods do not simultaneously consider item parameter estimates during scale transformation, meaning the parameters are not estimated concurrently. Haebara (1980) developed a method that simultaneously considers all parameter estimates, and later, Stocking and Lord (1983) proposed different methods similar to this approach. Both methods have been referred to as characteristic curve methods (Kolen & Brennan, 2014).

Once the parameters from different test forms have been placed on the same scale using these transformation methods, the IRT-based equating process can be applied. IRT-based equating methods include true score and observed score equating. The true score equating method, based on the mean of the conditional score distribution, assumes that for a given θ_i, the true scores τ_X (θ_i) and τ_Y (θ_i) are considered equivalent. The true score equivalent on Form Y for a given true score on Form X is defined as τ_X^{-1} corresponding to the true score τ_X (Kolen & Brennan, 2014).

The other equating method in IRT, known as the observed score equating method, utilizes the IRT model to estimate the distribution of the number of correctly answered items observed in each of the test forms to be equated. Subsequently, the equating process is conducted using equipercentile equating methods (Kolen & Brennan, 2014).

After discussing IRT equating methods, it is pertinent to briefly explain the local item independence assumption, which is one of the fundamental assumptions of IRT. Yen (1993) interpreted this assumption to mean that the true score or latent trait value contains all the information regarding the test taker's performance and that the contribution of each item in the test should be independent of the contributions of other items. This assumption is best understood within the context of IRT models, indicating that for a given value of the latent variable θ, the joint probability of correct responses to a pair of items is equal to the product of the probabilities of correct responses to the individual items. This relationship is mathematically represented in Equation 5 (Chen & Thissen, 1997).

$$p(U = U|\theta) = \prod_{i=1}^{I} p(u_i|\theta) = p(u_1|\theta)p(u_2|\theta) \dots p(u_1|\theta) \quad (5)$$

In summary, if the IRT model is correctly specified, item responses should be locally independent when the latent trait, referred to as theta (θ), is held constant (Yen, 1984). In other words, this implies that the test taker's response to any given item in the test does not influence their response to another item.

The violation of this assumption, which leads to biased estimates of person and item parameters, an overestimation of reliability, and biased calculations of equating errors (Chen & Thissen, 1997; Sireci *et al.*, 1991; Tuerlinckx & De Boeck, 2001), should not be overlooked. Some researchers (Sireci *et al*., 1991; Wainer, 1995; Wainer & Thissen, 1996; Wainer & Wang, 2000; Yen, 1993) have reported findings indicating that when there is dependence between items, the standard error estimation of measurement is low within the framework of classical test theory, and consequently, the reliability coefficient is estimated to be higher. The concern that applying unidimensional IRT-based test equating procedures to equate test scores, when the test forms are multidimensional, may lead to erroneous equating results has prompted researchers to develop Multidimensional Item Response Theory (MIRT)-based equating procedures. The objective of MIRT test equating is consistent with that of the IRT equating process; however, the latent trait universe is generally multidimensional rather than unidimensional. This complexity increases as the number of dimensions in the test rises, making it more challenging to position scores from scale linking or parallel forms within the same coordinate system (Peterson, 2014).

Lee and Brossman (2012) hypothesized that item types are the source of dimensionality in mixed-format tests and proposed an observed score equating method to equate these tests. In this method, responses to each item type (open-ended and multiple-choice) are associated with two abilities (θ_1, θ_2), and each latent trait is modeled within two unidimensional IRT frameworks. This equating method proposed by Lee and Brossman (2012) is referred to as the Simple Structured MIRT (SS-MIRT) observed score equating method, and the assumptions of this method are presented below:

a) The items in the test measure a characteristic corresponding to a particular item type and are associated with these characteristics.
b) Each item group can be modeled by a unidimensional IRT model.
c) Test takers responding to both the new and old forms are considered randomly equivalent with respect to θ_1 and θ_2.

To perform SS-MIRT observed score equating, items are first calibrated using the SS-MIRT model. θ_1 and θ_2 are calibrated separately for the X and Y forms. The parameter estimates for θ_1 items in the test forms are placed on the same θ_1 metric. Similarly, the parameter estimates for θ_2 items in the test forms to be equated are placed on the same $\theta_2$metric. Additionally, in

the randomized groups design, the correlation between the θ_1 and θ_2 traits is assumed to be equal for both groups.

Based on the item parameter estimates, conditional observed score distributions are obtained for each form across each latent trait dimension (θ_1 and θ_2). These distributions can be generated using an extended version of the Lord-Wingersky algorithm (Lee & Brossman, 2012). The total observed score is the sum of the weighted scores from different item types or content areas ($X = w_1X_1 + w_2X_2$).

To generate marginal total score distributions, the conditional total score distributions are summed over a bivariate latent trait distribution, g(θ_1 , θ_2), as in Equation 6:

$$f(x) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x|\theta_1,\theta_2)g(\theta_1,\theta_2)d\theta_1 d\theta_2 \quad (6)$$

The marginal observed score distributions are computed for both forms. Then, based on these distributions, the test forms are equated using equipercentile equating methods.

Equating methods based on CTT and IRT are critically important for ensuring the fair and comparable evaluation of tests. Numerous studies have compared the performance of these methods in terms of accuracy, stability, and bias, identifying which method is more effective under specific conditions. These studies offer significant findings to enhance the accuracy of test equating processes, contributing both theoretically and practically.

One of the studies in this field was conducted by Woodruff (1989), who analyzed and compared equating methods both analytically and empirically in scenarios where the covariance between test items and anchor items varied. The study highlighted that the Angoff-Levine method was more sensitive to the lack of content balance between test and anchor items compared to the Tucker method. Furthermore, Woodruff emphasized that the Congeneric-Levine method performed reasonably well in addressing unexpected situations in practice. However, he argued that the Tucker or Angoff-Levine methods could be more straightforward to apply than the Congeneric-Levine method, provided their assumptions were satisfied. Hanson *et al.* (1994) examined smoothing equipercentile equating (pre-smoothing and post-smoothing) and linear equating methods under a random groups design. They reported that both smoothing methods improved the estimation of the equipercentile equating function and observed that the two methods exhibited similar performance in terms of equating error. Tsai (1997) demonstrated that the equating error in traditional test equating methods decreased as the sample size increased. However, he also determined that equipercentile equating required larger sample sizes compared to linear and mean equating methods. Skaggs (2005) investigated the effectiveness of equating with small sample sizes and concluded that equating should not be performed when the sample size is 25 or fewer. As the sample size increases, the standard error of equating decreases, and bias shows minimal variation. These findings once again emphasize the importance of selecting the appropriate method and sample size in research applications. Studies conducted in Türkiye have also examined the effectiveness of equating methods based on CTT. Öztürk and Anıl (2012) observed that equipercentile equating was more appropriate than linear equating when comparing tests administered at different times. Demir and Güler (2014), in their study aimed at testing the statistical equivalence of different forms of concurrently administered tests, identified the Braun-Holland linear equating method as the most suitable. Tan (2015) compared polynomial and beta4 pre-smoothing methods with cubic spline post-smoothing methods under a single group design for equipercentile equating, finding that all three equating methods were effective. However, when considering the average bootstrap standard error and moment preservation criteria, they concluded that pre-smoothing methods resulted in fewer errors than cubic spline post-smoothing methods in small samples, with the beta4 pre-smoothing method being more suitable for use. Karagül (2020) observed that, in small sample groups, as sample size increased, equating error, bias, and RMSE values decreased for all equating methods compared. Furthermore, for sample sizes below 100, mean

equating produced fewer errors. These findings emphasize the importance of selecting the appropriate equating method, particularly for small sample sizes.

Karkee and Wright (2004) compared the performance of scale transformation methods and found that characteristic curve methods generated fewer errors than moment methods, with the Stocking-Lord method being the least error-prone. Kilmen (2010) emphasized that the Stocking-Lord scale transformation method produced fewer errors, while the mean-mean and mean-sigma methods resulted in more errors. Additionally, it was found that equating errors decreased as the sample size increased and when the ability levels of test participants were similar. Wang (2012) evaluated the performance of true and observed score equating methods based on factors such as test length, forms with different difficulty levels, ability distribution, parameter estimation methods, and test format. The results indicated that true and observed score equating produced very similar outcomes. Furthermore, for the same sample size, longer tests had fewer equating errors compared to shorter tests. Aksekioğlu (2017) found that the mean-sigma method was the least error-prone among the scale transformation methods. Regarding test equating methods, the author found that true and observed score equating provided similar results, with the observed score equating method producing fewer errors than the true score equating method.

Petersen *et al.* (1983) compared CTT and IRT-based equating methods and noted that for parallel tests, linear equating methods based on CTT were sufficient. However, when tests varied in terms of content and length, IRT methods provided more stable results. Han *et al.* (1997) compared the IRT true and observed score equating methods with the unsmoothing equipercentile equating method and found that the IRT true score equating method produced more stable results than both the IRT observed score and unsmoothing equipercentile equating methods. However, they also noted that the differences in equating stability among these three methods were minimal. Hagge (2010) emphasized that IRT methods, particularly in terms of bias and standard error, demonstrated better performance, with the IRT observed score equating method showing the lowest error values. Liu and Kolen (2011) found that IRT methods produced fewer errors in terms of bias and RMSE and were more robust to group differences. Powers *et al.* (2011) examined the sensitivity of the frequency estimation method, chained equipercentile equating, and the IRT true and observed score equating methods to group differences and found that these methods were not sensitive to group differences. They also noted that IRT equating methods produced less systematic equating error compared to traditional methods. Tanberkan Suna (2018) investigated the effects of violations of group invariance and found that, in addition to the Tucker and Braun-Holland equating methods, the mean-mean scale transformation method was more suitable for violations of group invariance. Moreover, the IRT observed score equating method was found to provide better results than the true score equating method. Mutluer (2021) compared CTT-based test equating methods, such as linear and equipercentile equating, and found that the equipercentile equating method produced fewer errors. Furthermore, among IRT scale transformation methods, the Stocking-Lord method provided better results, and when comparing the IRT true and observed score equating methods, the true score equating method produced fewer equating errors.

A review of the literature indicates that although there is an extensive body of research on test equating methods, studies specifically focusing on MIRT-based equating methods have only gained momentum in recent years. Consequently, research in this area remains relatively scarce. Lee and Brossman (2012) introduced the Simple-Structure Multidimensional Item Response Theory Observed Score Equating (SMO) method, founded on the premise that applying Unidimensional Item Response Theory (UIRT) equating methods to tests with a multidimensional data structure leads to inaccurate equating relationships. In their study, the researchers utilized multidimensional tests composed of items, each designed to measure a single proficiency. This research involved mixed-format tests incorporating both multiple-

choice and constructed-response items. Results from both real and simulated data demonstrated that the SMO method yielded satisfactory equating outcomes when the data were multidimensional. Moreover, the SMO method outperformed traditional UIRT equating methods. Subsequently, Kim (2018) developed the Simple-Structure Multidimensional IRT True Score Equating (SMT) method. Its performance was assessed using real, simulated, pseudo-form, and identical-form datasets. Additionally, the SMT method was compared with pre-smoothing equipercentile equating, IRT true score equating, IRT observed score equating, and the SMO method. Overall, the SMT method produced results comparable to existing methods. Notably, it provided more accurate equating outcomes than traditional IRT equating techniques. This superior performance was consistently observed across three studies involving diverse datasets and various evaluation criteria.

Brossman and Lee (2013) developed observed and true score equating methods within the framework of MIRT. They proposed three distinct procedures: the "Full MIRT Observed Score Equating Procedure", the "MIRT True Score Equating Procedure with a Unidimensional Approach," and the "MIRT Observed Score Equating Procedure with a Unidimensional Approach." In their study, they compared these methods to equipercentile equating and found that multidimensional IRT methods produced more consistent results. Lee *et al*. (2015) introduced a bifactor MIRT true score equating method, which demonstrated similar outcomes across various test models. Similarly, Lee and Lee (2016) assessed the applicability of bifactor MIRT observed score equating procedures for mixed-format tests, revealing results comparable to those obtained using UIRT. Tao and Cao (2016) expanded IRT equating methods under the testlet response model, exploring the effects of local item dependence. They concluded that observed score equating proved more advantageous when violations of local independence occurred.

Equating methods based on MIRT have gained increasing importance in psychometric literature and attracted significant attention from researchers. However, they have unfortunately not received adequate focus in Türkiye. A review of the literature reveals that studies addressing this area are limited in number. Atar and Yeşiltaş (2017) compared the performance of mean-mean, Stocking-Lord, and mean-standard deviation scale transformation methods adapted for multidimensional data. This comparison was based on factors such as ability distribution, sample size, percentage of common items in the test, and sample size. The study employed a common item pattern in non-equivalent groups. The results showed that, in terms of item difficulty and discrimination parameter estimates, the Stocking-Lord method demonstrated lower RMSE (Root Mean Square Error) and bias values compared to the other methods.

Gübeş-Öztürk (2019) examined the effect of multidimensionality on equated scores obtained through separate and simultaneous calibration structures, using simulation data under 40 conditions. The tests generated were assumed to measure two distinct abilities, θ_1 and θ_2. A common item pattern was used in non-equivalent groups. The results revealed that equating results obtained with separate calibration structures produced fewer equating errors and bias values compared to those obtained with simultaneous calibration methods. Additionally, when the degree of multidimensionality was significant, equating results obtained through simultaneous calibration exhibited the least random error.

Uğurlu (2020) investigated the relationship between test equating and differential item functioning (DIF) from a multidimensional perspective by exploring the population invariance property of equating. The study utilized simulated data generated under a simple-structured multidimensional IRT model. The researcher compared the simple-structured multidimensional IRT observed score, unidimensional IRT observed score, true score, and equipercentile equating methods under conditions involving form-differentiating DIF, form group ability mean differences, and interdimensional correlations based on population invariance values. The findings indicated that when the interdimensional correlation was .5, the simple-structured

multidimensional observed score equating method most accurately reflected the relationship between test equating and DIF. When the interdimensional correlation was .8 and .95, all methods yielded similar results, except for the equipercentile equating method in low-frequency scores.

1.1. Aim and Significance of the Study

The presence of a second dimension, resulting from a violation of the local item independence assumption, may lead to an inaccurate estimation of item parameters in the unidimensional IRT-based equating process (Chen, 2014). Such misestimations raise concerns about the accuracy of equating procedures (Kim *et al.*, 2020). If researchers overlook the violation of local item independence, the quality of IRT-based equating methods may be compromised (Chen, 2014).

To address these issues and improve the accuracy of equating relationships, researchers have developed MIRT-based equating methods. These equating methods account for different data structures including: "SS-MIRT Observed Score Equating" (W. Lee & Brossman, 2012), "Full-MIRT Observed Score Equating" (Brossman & Lee, 2013), "Unidimensional Approximation of MIRT True Score Equating Procedure" (Brossman & Lee, 2013), "Unidimensional Approximation of MIRT Observed Score Equating Procedure" (Brossman & Lee, 2013), "Bi-factor MIRT Observed Score Equating" (G. Lee & Lee, 2016), "Bi-factor MIRT True Score Equating" (G. Lee *et al*., 2015), "Testlet Model Theory- MIRT Observed Score Equating" (Tao & Cao, 2016), "Testlet Model Theory-MIRT True Score Equating (Tao & Cao, 2016), and "SS-MIRT True Score Equating" (Kim, 2018). This study focuses on the SS-MIRT observed score equating method, comparing its performance with traditional and IRT-based equating methods. The current study aims to contribute to the existing literature on SS-MIRT-based test equating and provide valuable insights for researchers and organizations conducting large-scale testing, highlighting the importance of selecting appropriate equating methods.

The aim of this study is to examine the effect of loading certain items onto a second dimension due to violations of the local independence assumption in unidimensional and dichotomous tests. The study also seeks to propose the most effective equating method by analyzing the impact of these violations on equating errors across different equating methods. Additionally, the study aims to interpret the unidimensionality effect. In line with these aims, the following research questions will be addressed:

1. How do the standard errors of equating, bias, and RMSE values obtained from traditional equating methods vary based on sample size, test length, and the proportion of items loaded onto the second dimension?
2. How do the standard errors, bias, and RMSE values in observed and true score equating compare based on sample size, test length, the proportion of items loaded onto the second dimension, and the scale transformation methods used in IRT equating?
3. How do the standard errors of equating, bias, and RMSE values in SS-MIRT-observed score equating vary with sample size, test length, and the proportion of items loaded onto the second dimension?
4. How do the RMSE values obtained from SS-MIRT observed score equating compare with those from traditional and IRT equating methods, particularly the methods that yield the lowest RMSE values, with respect to sample size, test length, and the proportion of items loaded onto the second dimension?

2. METHOD

2.1. Type of Research

The aim of this study was to compare the effects of violating the local independence assumption on various test equating methods for dichotomously scored tests under different conditions. To achieve this goal, data were generated for scenarios where the assumption of local item

independence was violated, with the aim of identifying the method that produces the least equating error. Given these characteristics, the study is categorized as a simulation study.

2.2. Research Design

A randomized group design was employed in this study. In this design, participants from the same population are randomly assigned to groups and receive different test forms (Cook & Eignor, 1991). The difference in performance between the groups indicates the variation in difficulty among the test forms. Therefore, it is essential to work with heterogeneous and large samples to minimize potential bias arising from the sample (Livingston, 2014).

2.3. Simulation Conditions

The simulation conditions included three main factors: test length (20, 40, 60), sample sizes (250, 1000, 5000), and the percentage of items loaded onto the second dimension (0%, 15%, 30%, 50%). As shown in Table 1, a total of 36 conditions (3 x 3 x 4) were considered, representing variations in sample sizes, test length, and the percentages of items loaded onto the second dimension.

Table 1. *Simulation conditions for the study.*

Factors	Conditions	Number of Conditions
Sample Size	250-1000-5000	3
Test Length	20-40-60	3
Percentage of Items Loaded onto the Second Dimension	%0-%15-%30-%50	4

2.4. Equating Methods to be Used

The study utilized several equating methods, including Classical Test Theory (mean equating, linear equating, equipercentile equating, and sixth-order polynomial loglinear smoothing equipercentile equating), Item Response Theory (true and observed score equating), and Multidimensional Item Response Theory (simple structured multidimensional observed score equating methods).

2.5. Data Generation

The R statistical software, version 4.1.1 (R Core Team, 2019), was used to generate the data for this study. The *mirt* package (Chalmers, 2012) was utilized to generate unidimensional data using the *for* loop and *simdata* command, while the *mirtCAT* package (Chalmers, 2016) was utilized to generate multidimensional data using the *for* loop and the *generate_pattern* command. For each dataset, 100 replications were conducted.

An important consideration for researchers conducting simulation studies is that the generated datasets must represent real responses (Way *et al.*, 1988). To meet this requirement, the parameters of the distributions were derived from the data of the PISA 2018 Mathematics Test Form-1 and were utilized in generating the study's data. Regarding the violation of the local independence assumption, the distributions of actual item parameters were obtained from the Türkiye sample of the PISA 2018 Mathematics Test, and these parameters were utilized for data generation. The PISA 2018 mathematics test data were initially analyzed within a unidimensional framework, with the fit indices presented in Table 2 and the resulting parameter values presented in Supplementary Material 1 (see Table S1).

Upon examining the fit indices (RMSEA < .05; CFI and TLI ≥ .95), it is evident that the data fit the unidimensional model well (Kline, 2015; Tabachnick & Fidell, 2013). Subsequently, the residual correlations between item pairs were examined, revealing residual correlations of .20 and above among certain item pairs (items 16, 17, and 18). Residual correlations of .20 and above indicate a violation of local independence for these items (Yen, 1993).

Table 2. *Model fit indices for unidimensional and SS-MIRT models based on the PISA 2018 Mathematics Test - Form 1 (Türkiye sample).*

Unidimensional IRT (2PLM)		SS-MIRT (2PLM)	
RMSEA	.033	RMSEA	.023
CFI	.955	CFI	.979
TLI	.949	TLI	.976
SRMR	.089	SRMR	.083

Accordingly, these three items, which were suspected to violate the local independence assumption, were loaded onto a second dimension, and the analyses were repeated. The fit indices from this subsequent analysis showed an improvement compared to those obtained from the unidimensional model. Based on these findings, it was concluded that these three items (items 16, 17 and 18) were indeed violating the local item independence assumption and thus were appropriately modeled as part of a second dimension. The parameters derived from this two-dimensional structure were subsequently used for data generation.

For unidimensional data, item discrimination parameters were drawn from a log-normal distribution with a mean of 0.53 and a standard deviation of 0.178, while item difficulty parameters were derived from a normal distribution with a mean of 0.327 and a standard deviation of 0.5. The individual θ parameters were obtained from a normal distribution with a mean of 0 and a standard deviation of 1. In cases where the local independence assumption was violated, the item discrimination parameters for items on the first dimension were sourced from the same log-normal distribution with a mean of 0.53 and a standard deviation of 0.178. Conversely, the discrimination parameters for items on the second dimension were obtained from a log-normal distribution with a mean of 0.70 and a standard deviation of 0.21, while the item difficulty parameter remained consistent, drawn from a normal distribution with a mean of 0.327 and a standard deviation of 0.5. Once the item parameters were established, ability parameters were generated using a bivariate normal distribution BN(0, 0, 0, 1, 1, 0.5). Data generation was conducted using 100 replications under the two-parameter logistic model. This study adopted a two-dimensional structure referred to as “simple structure” This term, first introduced by Thurstone (1947), indicates that each item loads on only one dimension, without cross-loading onto other dimensions (McDonald, 2000; Sass & Schmitt, 2010).

In the two-dimensional data sets, the correlation between dimensions was fixed at .5. Kahraman (2013) noted that when the correlation between latent traits is .70 or higher, the boundary between unidimensionality and multidimensionality becomes ambiguous. Furthermore, when the correlation between latent traits reaches .80 or above, unidimensional IRT models exhibit resistance to violations of the unidimensionality assumption (Kahraman & Kamata, 2004; Kahraman & Thompson, 2011). Multidimensional test equating studies have shown that when the correlation between latent abilities is low, equating errors increase (Lee & Brossman, 2012). Furthermore, a correlation of .5 between latent traits has been found to most accurately reflecting the equating relationships (Uğurlu, 2020). Based on these findings, the correlation between the dimensions in the two-dimensional data generated for this study was set at .5.

The residual correlations between item pairs in the generated datasets were examined, and it was observed that the residual correlations of items loading onto the second dimension, due to the violation of the local independence assumption, were greater than .20. Supplementary Material 2 (see Table S2) provides an example of the item residual correlation table for a dataset consisting of 20 items from 1,000 participants, where three items violate the local independence assumption.

Another aspect considered in the data generation process is as follows: Data were generated according to the 2PLM (Two-Parameter Logistic Model) based on the distribution mentioned

above, under the condition that 15% of the items violate the local independence assumption. In the case where 30% of the items in the test violate the local independence assumption and load onto the second dimension, the item parameters estimated under the first condition were fixed, and only the new item parameters that would load onto the second dimension were derived from the aforementioned distribution to create the new dataset.

2.6. SS-MIRT Equating Process

The SS-MIRT equating method can be applied to any number of item sets; however, in this study, it was applied to a two-dimensional test where items load onto a second dimension when the local independence assumption is violated. In this context, Forms X and Y represent the new and old forms, respectively. A unidimensional IRT model was assumed for each dimension. Although various combinations of unidimensional IRT models may be employed, a two-parameter logistic model (Lord, 1980) was preferred for each dimension in this study to better capture item discrimination and difficulty. In this study, items that violated the local independence assumption were loaded onto a second dimension, with a specified correlation between dimensions. Accordingly, this model is referred to as the "Simple Structure-MIRT Model". Composite scores were created by summing the scores from the two content areas, and for the purposes of this study, only raw scores were equated.

The SS-MIRT observed score equating methods follow these steps:

1. Based on the SS-MIRT model, item parameters for both forms were estimated on the same scale (Calibrate θ_1 and θ_2 separately for each form).
2. Conditional observed score distributions were obtained for each dimension from both forms.
3. Conditional total score distributions were obtained for each form using the conditional observed score distributions.
4. A bivariate (θ_1 and θ_2) normal ability distribution was created.
5. Marginal observed score distributions were derived for each form by summing the conditional total score distributions over the bivariate normal theta (θ) distribution.
6. Equipercentile equating was performed based on the two marginal total score distributions for Form X and Form Y.

2.7. Evaluation Criteria

In this study, the equating results were evaluated using Standard Error (SE), Bias, and Root Mean Square Error (RMSE) values. 'Bias' refers to the systematic error in the equating process, 'RMSE' indicates the overall error in equating, and 'SE' represents the standard error of equating. The mathematical expressions for these evaluation criteria are listed in Equations 7, 8, and 9.

$$Bias_i = \frac{1}{R} \sum_{r=1}^{R} (\hat{e}_y(x_i) - e_y(x_i)) \tag{7}$$

$$SE_i = \sqrt{\frac{\sum_{r=1}^{R} (\hat{e}_y(x_i) - \bar{e}_y(x_i))^2}{R}} \tag{8}$$

$$RMSE_i = \sqrt{Bias_i^2 + SE_i^2} \tag{9}$$

In the equations, R represents the number of replications (100); i refers a score point; x_i is the raw or scale score at point i; $\hat{e}_y(x_i)$ represents the equivalent score of the old form x_i in the new form; $e_y(x_i)$ represents the criterion equivalent score, and $\bar{e}_y(x_i)$ represents the average of the equivalents of the new form score x_i over R (100) replications.

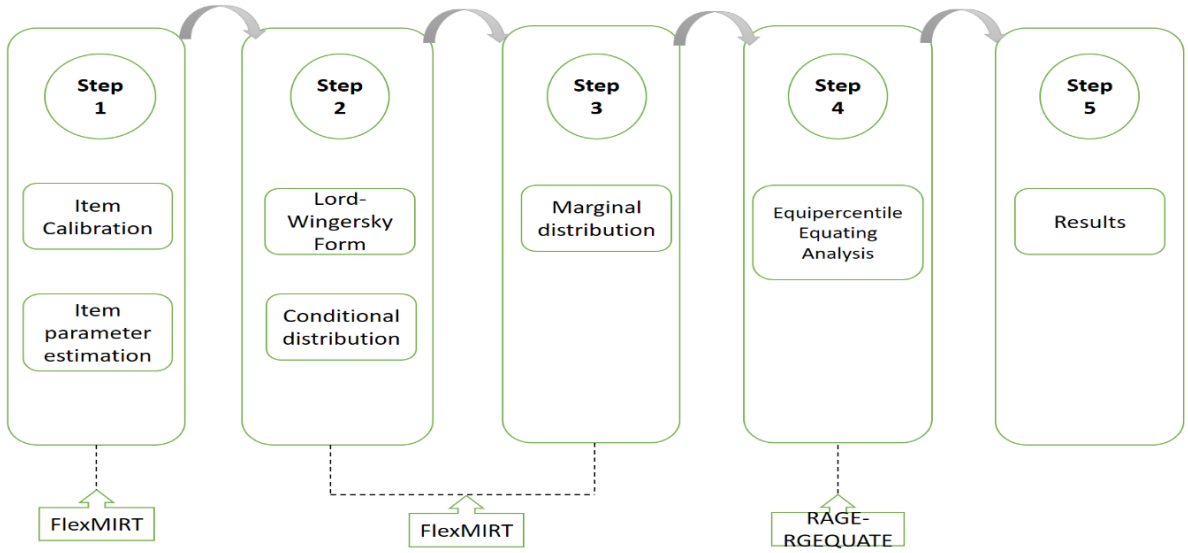
After calculating the Bias, SE, and RMSE values, these were scaled between 0 and 1 by dividing by the number of items in the test. This scaling was done to facilitate comparisons under varying

conditions of test length. Readers should note that the values presented in the study's tables are these weighted values.

2.8. Data Analysis

In this study, data generation was performed using the *mirt* package developed by Chalmers (2012) and the *mirtCAT* package developed by Chalmers (2016) in R statistical software (version 4.1.1). Detailed information regarding this process is provided in the data generation section. Traditional test equating analyses were conducted using the *equate* package (Albano, 2016) in R (version 4.1.1). Additionally, analyses for scale transformation methods in IRT-based observed score and true score equating were carried out using the *equateIRT* package (Battauz, 2015) in R (version 4.1.1). In this research, a total of 36 conditions were considered, representing variations in sample size, test length, and the percentages of items loaded onto the second dimension. For each condition, 100 replications were used, resulting in 100 Form Y and 100 Form X test datasets. These datasets were equated using traditional, observed and true score equating and SS-MIRT equating methods. This means that for each equating method, 100 equating processes were performed for each condition. After the equating process, the averages of the SE, Bias, and RMSE errors were calculated, and the obtained values were scaled between 0 and 1 by dividing them by the number of items in the test. This scaling was performed to facilitate the comparison of errors across different item count conditions.

Figure 1. *SS-MIRT equating process.*



To perform SS-MIRT observed score equating, items were first calibrated separately for θ_1 and θ_2 in both Form X and Form Y using the SS-MIRT model with flexMIRT software (Cai, 2020). Item parameters were estimated using simultaneous calibration. One of the main advantages of simultaneous calibration is that item calibration and scale linking are performed concurrently in a single process (Hanson & Béguin, 2002; Kim & Cohen, 1998). As a result, the parameter estimates for θ_1 items in Form X and Form Y were aligned on the same θ_1 proficiency scale, and the parameter estimates for θ_2 items were aligned on the same θ_2 proficiency scale. Based on these estimated item parameters, conditional observed score distributions for each form in each ability dimension (θ_1 and θ_2) were obtained using flexMIRT (These conditional distributions are obtained using the Lord and Wingersky formula). Subsequently, the marginal total score distribution was computed using the estimated bivariate proficiency distribution for the forms. The marginal observed score distributions for each form were obtained by summing the conditional total score distributions over the bivariate normal proficiency distribution. Finally, based on the obtained marginal total score distributions of Forms X and Y, an

equipercentile equating analysis was conducted using the RAGE-RGEQUATE software. Figure 1 summarizes the observed score equating process of the SS-MIRT based test equating (Kim, 2022) described above.

3. RESULTS

3.1. Results for Traditional Equating Methods

The standard error (SE) and bias values vary according to different sample sizes, test lengths, and equating methods. Supplementary Material 3 (see Table S3) presents the mean standard error values obtained from traditional equating methods. These data generally show a decrease in standard error values as sample size increases. However, the effect of the number of items varies depending on the equating method. For instance, with a sample size of 250, increasing the number of items from 20 to 40 results in a decrease in standard error values, while an increase in the number of items from 40 to 60 leads to an increase in the standard error values. Bias values, on the other hand, vary depending on the equating methods and test conditions. Supplementary Material 4 (see Table S4) presents the mean bias values calculated from traditional equating methods. Bias values change according to sample size, with some equating methods producing lower bias values for different sample sizes. For example, with a sample size of 250, bias values decrease as the number of items increases, whereas for sample sizes of 1000 and 5000, the differences in bias values between the equating methods become more pronounced.

Table 3 presents the mean RMSE values obtained from traditional equating methods. It was foundthat the linear equating method yields the lowest RMSE value, while for 40 and 60 items, the mean equating method produces the lowest RMSE value for a sample size of 250 with 20 items. For a sample size of 1000, the linear equating method consistently yields the lowest RMSE value across all examined conditions. However, for a sample size of 5000, the method that yields the lowest RMSE value varies depending on the conditions considered. In summary, there is no single equating method that consistently produces the lowest RMSE value across all conditions; rather, the methods yielding the lowest RMSE values differ based on specific conditions. Upon examining the effect of the violation of the local independence assumption on RMSE values, it is observed that there is no single equating method least affected by this violation. In summary, the impact of the violation of the local independence assumption on equating methods varies across all conditions.

3.2. Results for IRT Observed and True Score Equating

The mean standard error values for observed and true score equating were examined using the mean-standard deviation, mean-mean, Stocking-Lord, and Haebara scale transformation methods. It was found that as sample size and test length increased, standard error values decreased. Among the scale transformation methods, the Stocking-Lord method consistently yielded the lowest standard error values, while the mean-standard deviation method produced the highest. The Haebara and Stocking-Lord methods were the least affected by violations of the local independence assumption, with the Haebara method being slightly more robust. Furthermore, the observed score equating method produced lower standard error values compared to the true score equating method, although the difference between the two methods was small and further diminished as sample size and test length increased (see Supplementary Table S5).

Regarding bias values, the Stocking-Lord method consistently produced the lowest bias values across all conditions. It was found that the Stocking-Lord method was the least affected and the Haebara method was the most affected by violations of the local independence assumption. When comparing observed and true score equating methods, the bias values were generally similar, although the true score equating method produced lower bias values under certain conditions (see Supplementary Table S6).

Table 4 presents the RMSE means obtained using the scale transformation methods in observed and true score equating. An examination of Table 4 reveals that, with the exception of the Haebara method, the RMSE values in observed and true score equating performed using other methods tend to decrease as the sample size increases. In both observed and true-score equating, a sample size of 250 results in the highest RMSE values, whereas a sample size of 5,000 yields the lowest.

Although increasing the number of test items generally leads to a reduction in RMSE values, this pattern is disrupted under certain conditions. Nonetheless, it has been determined that for all scale transformation methods, the RMSE values calculated with 20 test items are higher than those calculated with 60 items. This finding suggests that as the number of test items increases, RMSE values decrease in both observed and true-score equating. Lastly, the Stocking-Lord scale transformation method was identified as the method with the lowest RMSE values in observed and true score equating, while the mean-standard deviation method exhibited the highest RMSE values. Furthermore, when assessing the impact of violating the local independence assumption on the RMSE values obtained from scale transformation methods in observed and true score equating, the Stocking-Lord method was found to be the least affected.

When comparing the RMSE values of both methods, it is observed that the values obtained from the observed score equating method are lower. Regarding the violation of the local independence assumption, the differences between the values calculated for unidimensional and multidimensional tests are generally similar. However, under the conditions of a sample size of 1,000 with 40 and 60 items in the test, the observed score equating method yields more robust results against the violation of the local independence assumption.

In conclusion, when comparing the standard error of measurement, bias, and RMSE values calculated under both methods, it has been determined that the observed score equating method demonstrates better performance in terms of having lower error values and being less affected by violations of the local independence assumption.

Table 3. *Mean RMSE values obtained from traditional equating methods.*

N	Item	Number of items in the test											
		20				40				60			
	Ratio	Mean	Linear	U-Eq	S-Eq	Mean	Linear	U-Eq	S-Eq	Mean	Linear	U-Eq	S-Eq
250	0%	.0510	.0495	.0556	.0560	.0450	.0511	.0521	.0523	.0277	.0518	.0359	.0373
	15%	.0652	.0621	.0640	.0644	.0363	.0414	.0429	.0443	.0295	.0500	.0355	.0374
	30%	.0379	.0300	.0404	.0405	.0257	.0309	.0336	.0347	.0240	.0456	.0308	.0328
	50%	.0471	.0387	.0458	.0460	.0231	.0277	.0304	.0315	.0238	.0452	.0308	.0323
1000	0%	.0179	.0153	.0232	.0234	.0225	.0200	.0248	.0250	.0128	.0122	.0175	.0176
	15%	.0230	.0198	.0255	.0260	.0241	.0216	.0248	.0248	.0173	.0168	.0203	.0203
	30%	.0154	.0094	.0179	.0180	.0166	.0137	.0186	.0187	.0185	.0185	.0218	.0216
	50%	.0125	.0051	.0169	.0167	.0138	.0103	.0169	.0171	.0114	.0110	.0161	.0163
5000	0%	.0344	.0339	.0339	.0340	.0196	.0195	.0204	.0204	.0111	.0100	.0120	.0122
	15%	.0412	.0410	.0393	.0394	.0091	.0079	.0108	.0107	.0099	.0105	.0114	.0115
	30%	.0247	.0249	.0240	.0241	.0119	.0114	.0130	.0130	.0097	.0104	.0111	.0113
	50%	.0199	.0201	.0207	.0208	.0292	.0298	.0282	.0282	.0070	.0066	.0088	.0090

Note.: U-Eq = Unsmoothed Equipercentile, S-Eq = Smoothed Equipercentile. Item Ratio =Number of items loaded onto the second dimension

Table 4. *RMSE values obtained in observed and true score equating using scale transformation methods.*

			Number of items in the test											
	N	Item	20				40				60			
		Ratio	M-M	M-S	H	S.L	M-M	M-S	H	S.L	M-M	M-S	H	S.L
Observed Score Equating	250	0%	0.0055	0.0145	0.0051	0.0022	0.0037	0.0127	0.0034	0.0015	0.0031	0.0097	0.0032	0.0013
		15%	0.0129	0.0245	0.0053	0.0027	0.0054	0.0143	0.0050	0.0019	0.0041	0.0119	0.0032	0.0015
		30%	0.0142	0.0231	0.0048	0.0031	0.0055	0.0137	0.0072	0.0023	0.0036	0.0114	0.0033	0.0015
		50%	0.0078	0.0181	0.0107	0.0029	0.0042	0.0142	0.0036	0.0017	0.0041	0.0118	0.0053	0.0017
	1000	0%	0.0024	0.0077	0.0029	0.0010	0.0022	0.0062	0.0018	0.0008	0.0020	0.0049	0.0019	0.0008
		15%	0.0040	0.0127	0.0026	0.0012	0.0026	0.0082	0.0018	0.0009	0.0026	0.0060	0.0024	0.0009
		30%	0.0062	0.0123	0.0039	0.0018	0.0025	0.0069	0.0024	0.0010	0.0027	0.0058	0.0034	0.0010
		50%	0.0022	0.0085	0.0024	0.0009	0.0025	0.0065	0.0032	0.0012	0.0017	0.0059	0.0032	0.0008
	5000	0%	0.0010	0.0035	0.0026	0.0004	0.0011	0.0026	0.0016	0.0006	0.0008	0.0021	0.0025	0.0003
		15%	0.0018	0.0044	0.0013	0.0005	0.0031	0.0038	0.0022	0.0010	0.0011	0.0027	0.0019	0.0003
		30%	0.0036	0.0050	0.0029	0.0013	0.0023	0.0033	0.0027	0.0010	0.0012	0.0025	0.0018	0.0003
		50%	0.0017	0.0044	0.0042	0.0009	0.0017	0.0037	0.0021	0.0009	0.0012	0.0026	0.0037	0.0004
True Score Equating	250	0%	0.0057	0.0154	0.0052	0.0025	0.0038	0.0262	0.0035	0.0016	0.0031	0.0099	0.0032	0.0014
		15%	0.0128	0.0255	0.0056	0.0033	0.0055	0.0291	0.0051	0.0020	0.0042	0.0121	0.0031	0.0016
		30%	0.0138	0.0241	0.0054	0.0039	0.0057	0.0280	0.0071	0.0024	0.0037	0.0116	0.0032	0.0016
		50%	0.0079	0.0192	0.0108	0.0033	0.0043	0.0293	0.0038	0.0019	0.0041	0.0120	0.0054	0.0018
	1000	0%	0.0025	0.0080	0.0029	0.0011	0.0022	0.0127	0.0019	0.0008	0.0020	0.0049	0.0018	0.0008
		15%	0.0040	0.0132	0.0028	0.0013	0.0026	0.0169	0.0018	0.0010	0.0027	0.0059	0.0023	0.0009
		30%	0.0062	0.0129	0.0043	0.0019	0.0026	0.0142	0.0023	0.0011	0.0027	0.0057	0.0032	0.0011
		50%	0.0023	0.0090	0.0026	0.0010	0.0026	0.0133	0.0033	0.0012	0.0017	0.0060	0.0032	0.0008
	5000	0%	0.0010	0.0037	0.0030	0.0005	0.0012	0.0054	0.0015	0.0006	0.0008	0.0021	0.0025	0.0003
		15%	0.0018	0.0048	0.0015	0.0006	0.0031	0.0081	0.0024	0.0010	0.0011	0.0027	0.0018	0.0003
		30%	0.0036	0.0055	0.0026	0.0012	0.0021	0.0073	0.0029	0.0010	0.0011	0.0025	0.0016	0.0003
		50%	0.0018	0.0047	0.0043	0.0010	0.0016	0.0074	0.0021	0.0009	0.0012	0.0027	0.0035	0.0004

Note. M-M = Mean-Mean, M-S = Mean-Sigma, H = Haebara, S.L = Stocking and Lord, Item Ratio = Number of items loaded onto the second dimension

3.3. Results for SS-MIRT Observed Score Equating

The mean SE, Bias, and RMSE values obtained from the SS-MIRT observed score equating method are presented in Table 5. Unlike the other tables, Table 5 does not include the condition where the percentage of items loading on the second dimension is 0%. This omission is due to the mathematical structure of the SS-MIRT equating method, both test forms being equated must have a multidimensional structure.

Table 5. *SE, Bias, and RMSE values obtained in SS-MIRT observed score equating.*

N	Item Ratio	Number of items in the test								
		20			40			60		
		SE	BIAS	RMSE	SE	BIAS	RMSE	SE	BIAS	RMSE
250	15%	0.0016	0.0005	0.0017	0.0010	0.0003	0.0010	0.0008	-0.0003	0.0009
	30%	0.0015	0.0010	0.0018	0.0012	-0.0008	0.0014	0.0009	-0.0007	0.0011
	50%	0.0017	-0.0008	0.0019	0.0010	0.0006	0.0012	0.0008	0.0009	0.0012
1000	15%	0.0009	0.0007	0.0011	0.0005	0.0004	0.0006	0.0004	-0.0005	0.0006
	30%	0.0007	0.0009	0.0011	0.0005	0.0002	0.0005	0.0004	-0.0004	0.0006
	50%	0.0005	0.0005	0.0007	0.0004	-0.0003	0.0005	0.0003	0.0002	0.0004
5000	15%	0.0003	-0.0004	0.0005	0.0002	0.0005	0.0005	0.0001	0.0001	0.0001
	30%	0.0003	-0.0007	0.0008	0.0002	0.0003	0.0004	0.0001	0.0001	0.0001
	50%	0.0002	0.0005	0.0005	0.0001	-0.0004	0.0004	0.0001	0.0002	0.0002

Note. Item Ratio =Number of items loaded onto the second dimension

As the number of items in the test and the sample size increase, a reduction in the standard error of the equating has been observed. The sample size with the highest standard error is 250, while the sample size with the lowest standard error is 5000. In terms of the number of items in the test, the highest standard error values were found under the 20-item condition, and the lowest values were found under the 60-item condition. This finding supports the inference made earlier. Another notable finding is that as the sample size increases, the variation in standard error values across different amounts of items loading on the second dimension decreases. This can be interpreted as the effect of violating the local independence assumption diminishing with increasing sample size in terms of the standard error of equating. When examining the bias values obtained from the equating results, no systematic findings were observed regarding the amount of items loading on the second dimension, sample size, and the number of items in the test. The highest bias values were found under the condition of a sample size of 250 and 20 items. Looking at the RMSE values, it is evident that as the sample size and the number of items in the test increase, the error values decrease. The lowest RMSE values were obtained under the conditions of a sample size of 5000 and 60 items.

3.4. Comparison of SS-MIRT Observed Score Equating with Other Methods

When considering the RMSE values from traditional equating methods, it was found that there is no single equating method that consistently produces the least error or is the least affected by the violation of the local independence assumption across all analyzed conditions; rather, this varies depending on the examined conditions. For the sake of completeness in the research, the RMSE errors obtained from traditional equating methods are presented. Additionally, the values from the Observed Score Equating method, which yields the lowest RMSE and is the least affected by the violation of the local independence assumption within IRT, along with the RMSE values obtained from the SS-MIRT observed score equating method, are provided in Table 6.

Table 6. *RMSE values obtained from traditional equating methods, IRT observed score equating, and SS-MIRT observed score equating methods.*

		Number of items in the test																	
N	Item Ratio	20						40						60					
		Traditional Equating Methods				IRT	SS-MIRT	Traditional Equating Methods				IRT	SS-MIRT	Traditional Equating Methods				IRT	SS-MIRT
		Mean	Linear	U-Eq	S-Eq	Obs.	Obs.	Mean	Linear	U-Eq	S-Eq	Obs.	Obs.	Mean	Linear	U-Eq	S-Eq	Obs.	Obs.
250	15%	0.0652	0.0621	0.0640	0.0644	0.0027	0.0017	0.0363	0.0414	0.0429	0.0443	0.0019	0.0010	0.0295	0.0500	0.0355	0.0374	0.0015	0.0009
	30%	0.0379	0.0300	0.0404	0.0405	0.0031	0.0018	0.0257	0.0309	0.0336	0.0347	0.0023	0.0014	0.0240	0.0456	0.0308	0.0328	0.0015	0.0011
	50%	0.0471	0.0387	0.0458	0.0460	0.0029	0.0019	0.0231	0.0277	0.0304	0.0315	0.0017	0.0012	0.0238	0.0452	0.0308	0.0323	0.0017	0.0012
1000	15%	0.0230	0.0198	0.0255	0.0260	0.0012	0.0011	0.0241	0.0216	0.0248	0.0248	0.0009	0.0006	0.0173	0.0168	0.0203	0.0203	0.0009	0.0006
	30%	0.0154	0.0094	0.0179	0.0180	0.0018	0.0011	0.0166	0.0137	0.0186	0.0187	0.0010	0.0005	0.0185	0.0185	0.0218	0.0216	0.0010	0.0006
	50%	0.0125	0.0051	0.0169	0.0167	0.0009	0.0007	0.0138	0.0103	0.0169	0.0171	0.0012	0.0005	0.0114	0.0110	0.0161	0.0163	0.0008	0.0004
5000	15%	0.0412	0.0410	0.0393	0.0394	0.0005	0.0005	0.0091	0.0079	0.0108	0.0107	0.0010	0.0005	0.0099	0.0105	0.0114	0.0115	0.0003	0.0001
	30%	0.0247	0.0249	0.0240	0.0241	0.0013	0.0008	0.0119	0.0114	0.0130	0.0130	0.0010	0.0004	0.0097	0.0104	0.0111	0.0113	0.0003	0.0001
	50%	0.0199	0.0201	0.0207	0.0208	0.0009	0.0005	0.0292	0.0298	0.0282	0.0282	0.0009	0.0004	0.0070	0.0066	0.0088	0.0090	0.0004	0.0002

Note. U-Eq = Unsmoothed Equipercentile, S-Eq = Smoothed Equipercentile, IRT Obs. = IRT Observed Score Equating, SS-MIRT Obs. = Simple-structure MIRT observed score equating

Among the compared methods, it is clearly observed that the methods with the highest RMSE values are the traditional equating methods, while the method with the lowest RMSE value is the SS-MIRT observed score equating method. The RMSE values obtained from the IRT observed score equating method and the SS-MIRT observed score equating method are quite similar, whereas the RMSE values obtained from the traditional equating methods are significantly higher than those of the two observed score equating methods. In terms of the amount of items loading onto the second dimension, the method with the greatest variation in RMSE values is the traditional equating method, while in IRT-based methods (Unidimensional IRT and SS-MIRT), the variation is found to be less. It has also been found that the least variation occurs with the SS-MIRT observed score equating method. Based on these findings, it can be inferred that the SS-MIRT observed score equating method is less affected by the violation of the local independence assumption compared to the other equating methods. Another important finding is that as the number of items in the test and the sample size increase, the impact of the violation of the local independence assumption on the RMSE values obtained from all equating methods decreases. In light of these findings, it can be concluded that the SS-MIRT observed score equating method produces the lowest RMSE values and is more resilient to the violation of the local independence assumption.

4. DISCUSSION and CONCLUSION

The primary objective of this study is to examine the effects of certain items loading onto a second dimension due to the violation of the local independence assumption in unidimensional and dichotomously scored tests, using different test equating methods, and to propose the method that demonstrates the best performance. To this end, the test forms for equating have been constructed on a simple multidimensional structure, and the equating process was carried out. In this research, the source of the tests' multidimensionality is addressed as the violation of the local independence assumption. When item responses violate the local independence assumption in a unidimensional model, these items load onto another dimension (DeMars, 2010), and the presence of this second dimension leads to the misestimation of item parameters in the unidimensional IRT-based equating process (Chen, 2014). In this case, the accuracy of the equating relationships established as a result of the equating process is compromised. Although unidimensionality and local item independence are assumptions of IRT, research has shown that when there is dependence among items, the standard error of measurement is estimated to be lower in classical test theory (CTT) as well (Sireci et al., 1991; Wainer, 1995; Wainer & Thissen, 1996; Wainer & Wang, 2000; Yen, 1993).

In terms of sample size, when examining the equating methods included in the study, it was found that as sample size increases, the standard errors of the obtained equating decrease. According to the literature, Kolen and Brennan (2014) suggested that sample size affects equating errors. Additionally, Harris and Crouse (1993), Kilmen (2010), Kim (2018), Kim and Cohen (2002), Lee and Ban (2010), and Salmaner Doğan and Tan (2022) have demonstrated in their studies that equating errors decrease as sample size increases. Furthermore, Livingston (1993), Livingston and Kim (2009), and Skaggs (2005) proposed that a small sample size in the equating process could jeopardize the accuracy of the estimates, potentially increasing the standard errors of the equating. In this study, it was found that the equating method with the highest standard error was obtained with a sample size of 250, which appears to support the findings of the aforementioned studies. Another noteworthy finding regarding sample size is that, in this study, the standard error values obtained in the equating for the largest sample size of 5000 are closer to the standard error values obtained when the test forms are unidimensional, in comparison to the other sample sizes. Based on these findings, it can be concluded that, for all the equating methods compared, as the sample size increases, the effect of the violation of the local independence assumption on the standard errors of the equating decreases. When examining the bias values, no systematic findings were identified among the equating methods compared. However, the highest bias error value was observed with a sample size of 250.

Regarding the RMSE values, it was found that only under the condition of 60 items did the RMSE values decrease as the sample size increased. In IRT-based equating methods, the RMSE values obtained from the observed and true score equating methods, with the exception of the Haebara scale transformation method, generally showed a decreasing trend as the sample size increased. Similarly, in the SS-MIRT equating method, the RMSE values also decreased as the sample size increased. Based on these findings, it can be observed that using a larger sample size in the study reduces variability. This result is consistent with the findings of Kim *et al.* (2020). In test equating conducted under a random groups design, the reduction in standard errors of equating as the sample size increases can be attributed to the selected sample's better representation of the population. In other words, when working with larger samples, the sample distribution is increasingly likely to resemble the population distribution. As the sample size grows, the sample distribution becomes more normal and symmetric compared to smaller samples. This leads to improved test equating performance and more reliable standard error estimates. Furthermore, the increase in sample size results in more accurate item parameter estimates. A more precise estimation of item parameters, in turn, reduces the standard error of equating. In summary, when equating is performed with larger samples, it is believed that the sample better represents the population, item parameters are estimated more accurately, and this contributes to smaller standard errors. Li and Lissitz (2004) found in their study that sample size is a significant factor in the standard error of parameter estimates, with larger sample sizes leading to smaller standard errors. Zhang (2010) theoretically argued that, all else being equal, as the sample size increases, item parameter estimates would have smaller standard errors. The increase in random errors in test equating with smaller samples can be attributed to unstable parameter estimates. Zhang emphasized that stable and accurate parameter estimates for items are crucial for improving the accuracy of equating. These findings support the conclusions drawn from our study.

When examining the results in terms of the number of items in the test, it was generally found that, with the exception of the linear equating method, as the number of items in the test increased, the standard errors of equating decreased for the other equating methods. However, in the case of the linear equating method, it was concluded that as the number of items in the test increased, the standard errors of equating increased. This finding is consistent with the research of Aşiret (2014) and Çörtük (2022). In IRT true and observed scores, a decrease in standard errors of equating was generally observed as the number of items in the test increased, both with scale transformation methods loaded on the second dimension and with SS-MIRT observed score equating methods. The findings of our study are similar to those of Akour (2006), Gök and Kelecioğlu (2014), Lee *et al*. (2014), Kumlu (2019), and Wang *et al*. (2020).

When examining the standard errors of equating methods based on CTT, it is observed that, except for the sample size of 250, the linear equating method produced a lower standard error of equating and was less affected by the impact of the local independence assumption. However, at a sample size of 250, as the number of items in the test increased, the linear equating method was found to be both the method producing the highest error value and the one most affected by the violation of the local independence assumption. When examining bias error values, it was found that there is no single equating method that consistently has the least bias error value across all conditions. A similar result is observed in RMSE values, where no single equating method consistently produces the least RMSE value; rather, this varies according to the conditions. The inability to obtain a similar finding in the RMSE values for the linear equating method at sample sizes of 1000 and 5000, despite achieving the least standard error of equating, is due to the increase in bias values. In conclusion, the results obtained from CTT-based equating methods varied according to the types of errors and the conditions considered.

When examining SE, bias, and RMSE values for scale transformation methods within the context of IRT, the Stocking-Lord scale transformation method stands out as the one producing the least error values and being least affected by violations of the local independence

assumption compared to other scale transformation methods. In terms of SE, Bias, and RMSE values, when comparing true and observed score equating methods based on IRT, it is observed that, contrary to the values obtained from CTT-based equating methods, the values obtained from true and observed score equating methods are quite close to each other, with the differences between the two equating methods being very small. When examining the standard error values obtained from true and observed score equating methods, it was determined that the standard error values obtained from the observed score equating method were slightly smaller than those obtained from the true score equating method. However, in terms of bias values, the opposite finding was observed, with the true score equating method yielding lower error values. When examining the RMSE values, it is observed that the method producing the least error values and least affected by violations of the local independence assumption is the observed score equating method. Based on RMSE values, it can be concluded that the observed score equating method is more resilient to violations of the local independence assumption. The other equating method used in the study, the SS-MIRT observed score equating method, was compared with CTT-based equating methods and the IRT observed score equating method in terms of RMSE values. The comparison revealed that the SS-MIRT observed score equating method has lower RMSE values compared to the other methods. It was found that the method with the greatest variation in RMSE values based on the proportion of items loading on the second dimension is the traditional equating methods, whereas the variation is less in IRT-based equating methods. Additionally, the least variation was found under the SS-MIRT observed score equating method. Based on these findings, it can be concluded that when both test forms are multidimensional, the SS-MIRT equating method is less affected by violations of the local independence assumption.

Upon reviewing the literature, it has been observed that studies comparing MIRT-based equating methods with other equating methods have yielded more accurate results from MIRT methods (Choi, 2019; Kim, 2018; Kim *et al.*, 2020; Lee & Brossman, 2012; Lee & Lee, 2016; Lee *et al*., 2014; Peterson & Lee, 2014; Tao & Cao, 2016). The results obtained from our research align with these findings. Lee and Brossman (2012) argued that in their studies, where they took the equipercentile equating method as a reference value, the results of the SS-MIRT observed score equating method were more similar to those of the equipercentile equating method. According to the researchers, this is because the test forms to be equated are not unidimensional, which may cause IRT equating methods to be more affected by violations of this assumption. They suggested that the SS-MIRT observed score equating method, by taking multidimensionality into account and due to the lack of an assumption regarding unidimensionality in the equipercentile equating method, is less affected by the violation of this assumption compared to IRT equating methods. In this research, when comparing the results obtained from CTT-based equating methods with those obtained from IRT equating methods, it was found that, contrary to the aforementioned study, CTT-based equating methods were more affected by the violation of the local independence assumption in cases where the test forms were multidimensional. This aspect distinguishes our study from the previously mentioned research. While Lee *et al.* (2014) noted that IRT-based equating methods performed better than equipercentile equating methods, Peterson and Lee (2014) indicated that MIRT and IRT methods yielded similar results, but the equipercentile equating method produced more divergent results. In this regard, the findings of these two studies are consistent with the results of our research.

Like any scientific study, this research has certain limitations. First, the research results were obtained from simulation datasets, and the findings are constrained by the simulation conditions. Additionally, it is always beneficial to conduct similar studies using real datasets and compare the results with those of this research. Another significant limitation of this study is the lack of an absolute criterion to assess the accuracy and precision of the compared equating methods. A review of the literature shows that many studies use different evaluation criteria to

assess equating methods. Finally, since this research is a simulation study, the generated data were produced within the framework of IRT and MIRT. This may have provided an advantage for IRT and MIRT-based equating methods compared to traditional equating methods. This concern can be addressed through equating studies conducted with real datasets.

Based on the results presented above, it is an undeniable reality that the dimensional structure of the data must be meticulously examined when conducting equating. Many tests, especially educational and psychological assessments, are inherently multidimensional. Moreover, this multidimensionality can have various sources. While the theoretical assumption that a test participant needs only a single latent trait to answer an item correctly may seem ideal, in practice, the cognitive processes of the test participants can be somewhat more complex. If any evidence of multidimensionality is obtained as a result of the analyses conducted, the preference for multidimensional IRT equating methods, as indicated by the findings of this research, will enable more accurate equating results to be achieved.

Acknowledgments

This study is a part of the doctoral dissertation of the first author under the supervision of the second author.

Declaration of Conflicting Interests and Ethics

The authors declare no conflict of interest. This research study complies with research publishing ethics. The scientific and legal responsibility for manuscripts published in IJATE belongs to the authors.

Contribution of Authors

Mehmet Fatih Doğuyurt: Literature review, Resources, Methodology, Data analysis, Reporting and Writing-original draft. **Şeref Tan:** Supervision.

Orcid

Mehmet Fatih Doğuyurt https://orcid.org/0000-0001-9206-3321
Şeref Tan https://orcid.org/0000-0002-9892-3369

REFERENCES

Aiken, L.R. (2000). *Psychological testing and assesment* (10th ed.). Allyn and Bacon.

Aksekioğlu, B. (2017). *Madde tepki kuramına dayalı test eşitleme yöntemlerinin karşılaştırılması: PISA 2012 fen testi örneği [Comparison of test equating methods based on item response theory: PISA 2012 science test sample]*. [Master's Thesis, Akdeniz University]. Higher Education Institution National Thesis Center.

Akour, M.M.M. (2006). *A comparison of various equipercentile and kernel equating methods under the random groups design.* [Doctoral Dissertation, Graduate College of The University of Iowa]. ProQuest Dissertations and Theses Global.

Albano, A.D. (2016). equate: An R package for observed-score linking and equating. *Journal of Statistical Software, 74*, 1-36. https://doi.org/10.18637/jss.v074.i08

Angoff, W.H. (1984). *Scales, norms, and equivalent scores*. Educational Testing Service.

Aşiret, S. (2014). *Küçük Örneklemlerde test eşitleme yöntemlerinin çeşitli faktörlere göre incelenmesi [Factors affecting the test equating method using small samples]*. [Master's Thesis, Mersin University]. Higher Education Institution National Thesis Center.

Atar, B., & Yeşiltaş, G. (2017) Çok boyutlu eşitleme yöntemlerinin eşdeğer olmayan gruplarda ortak madde deseni için performanslarının incelenmesi [Investigation of the performance of multidimensional equating procedures for common-item nonequivalent groups design]. *Journal of Measurement and Evaluation in Education and Psychology, 8*(4), 421-434. https://doi.org/10.21031/epod.335284

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.

Battauz, M. (2015). equateIRT: An R package for IRT test equating. *Journal of Statistical Software, 68*, 1-22. https://doi.org/10.18637/jss.v068.i07

Brossman, B.G., & Lee, W.C. (2013). Observed score and true score equating procedures for multidimensional item response theory. *Applied Psychological Measurement, 37(6),* 460-481. https://doi.org/10.1177/0146621613484083

Chalmers, R.P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of statistical Software, 48*, 1-29. https://doi.org/10.18637/jss.v048.i06

Chalmers, R.P. (2016). mirtCAT: Computerized adaptive testing with multidimensional item response theory. *Journal of statistical Software, 71*(5), 1-38. https://doi.org/10.18637/jss.v071.i05

Chen, J. (2014). *Model selection for IRT equating of testlet-based tests in the random groups design.* [Doctoral Dissertation, Graduate College of The University of Iowa]. ProQuest Dissertations and Theses Global.

Chen, W.H., & Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. *Journal of Educational and Behavioral Statistics*, *22*(3), 265-289. https://doi.org/10.1177/0265532220927487

Choi, J. (2019). *Comparison of MIRT observed score equating methods under the common-item nonequivalent groups design*. [Doctoral Dissertation, Graduate College of The University of Iowa]. ProQuest Dissertations and Theses Global.

Cook, L.L., & Eignor, D.R. (1991). An NCME module on IRT Equating methods. *Educational Measurement: Issues and Practice, 10*(3), 191-199. https://doi.org/10.1111/j.1745-3992.1991.tb00207.x

Çörtük, M. (2022). *Çok kategorili puanlanan maddelerden oluşan testlerde klasik test kuramı ve madde tepki kuramına dayalı test eşitleme yöntemlerinin karşılaştırılması [Comparison of test equating methods based on classical test theory and item response theory in polytomously scored tests]*. [Master's Thesis, Akdeniz University]. Higher Education Institution National Thesis Center.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory.* Harcourt Brace Javonich College.

Cui, Z. (2006). *Two new alternative smoothing methods in equating: The cubic B-spline presmoothing method and the direct presmoothing method*. [Doctoral dissertation, Graduate College of The University of Iowa]. ProQuest Dissertations & Theses Global.

De Gruijter, D.N., & Leo, J.T. (2007). *Statistical test theory for the behavioral sciences*. Chapman and Hall/CRC.

DeMars, C. (2010). *Item response theory.* Oxford University.

Demir, S., & Güler, N. (2014). Ortak maddeli denk olmayan gruplar desenine ilişkin test eşitleme çalışması [Study of test equating on the common item nonequivalent group design]. *International Journal of Human Sciences, 11*(2), 190-208.

Donlon, T.F. (1984). *The College Board technical handbook for the scholastic aptitude test and achievement tests.* College Entrance Examination Board.

Finch, H., French, B.F., & Immekus, J.C. (2014). *Applied psychometrics using SAS*. IAP.

Gök, B. , & Kelecioğlu, H. (2014). Denk olmayan gruplarda ortak madde deseni kullanılarak madde tepki kuramına dayalı eşitleme yöntemlerinin karşılaştırılması [Comparison of IRT equating methods using the common-item nonequivalent groups design]. *Mersin University Journal of the Faculty of Education, 10*(1), 120-136

Gübeş, N.Ö. (2019). Test eşitlemede çok boyutluluğun eş zamanlı ve ayrı kalibrasyona etkisi [The effect of multidimensionality on concurrent and separate calibration in test equating].

Hacettepe University Journal of Education, *34*(4), 1061-1074. https://doi.org/10.16986/HUJE.2019049186

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

Hagge, S.L. (2010). *The impact of equating method and format representation of common items on the adequacy of mixed-format test equating using nonequivalent groups*. [Doctoral Dissertation, Graduate College of The University of Iowa]. ProQuest Dissertations and Theses Global.

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

Han, T., Kolen, M., & Pohlmann, J. (1997). A comparison among IRT true-and observed-score equatings and traditional equipercentile equating. *Applied Measurement in Education, 10*(2), 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.

Hanson, B.A., Zeng, L., & Colton, D.A. (1994). *A comparison of presmoothing and postsmoothing methods in equipercentile equating* (No. 94). American College Testing Program.

Harris, D.J., & Crouse, J.D. (1993). A study of criteria used in equating. *Applied Measurement in Education, 6* (3),195-240. https://doi.org/10.1207/s15324818ame0603_3

Kahraman, N. (2013). Unidimensional interpretations for multidimensional test items. *Journal of Educational Measurement, 50*(2), 227-246. https://doi.org/10.1111/jedm.12012

Kahraman, N., & Kamata, A. (2004). Increasing the precision of subscale scores by usingout-of-scale information. *Applied Psychological Measurement, 28,* 407-426. https://doi.org/10.1177/0146621604268736

Kahraman, N., & Thompson, T. (2011). Relating unidimensional IRT parameters to a multidimensional response space: A review of two alternative projection IRT models for subscalescores. *Journal of Educational Measurement, 48,* 146-164. https://doi.org/10.1111/j.1745-3984.2011.00138.x

Karagül, A.E. (2020). *Küçük örneklemlerde çok kategorili puanlanan maddelerden oluşan testlerde klasik test eşitleme yöntemlerinin karşılaştırılması [Comparison of classical test equating methods with polytomously scored tests and small samples].* [Master's thesis, Ankara University]. Higher Education Institution National Thesis Center.

Karkee, T.B., & Wright, K.R. (2004, April). *Evaluation of linking methods for placing threeparameter logistic ıtem parameter estimates onto a one-parameter scale*. Paper presented at the Annual Meeting of the American Educational Research Association in San Diego, California.

Kilmen, S. (2010). *Comparison of equating errors estimated from test equation methods based on ıtem response theory according to the sample size and ability distribution*. [Doctoral Dissertation, Ankara University]. Higher Education Institution National Thesis Center.

Kim, S.Y. (2018). *Simple structure MIRT equating for multidimensional tests.* [Doctoral Dissertation, Graduate College of The University of Iowa]. ProQuest Dissertations & Theses Global.

Kim, S.H., & Cohen, A.S. (1998). Detection of differential item functioning under the graded response model with the likelihood ratio test. *Applied Psychological Measurement, 22*(4), 345-355.

Kim, S.Y., Lee, W.C., & Kolen, M.J. (2020). Simple-structure multidimensional item response theory equating for multidimensional tests. *Educational and Psychological Measurement, 80*(1), 91-125. https://doi.org/10.1177/0013164419854208

Kline, R.B. (2015). *Principles and practice of structural equation modeling.* Guilford.

Kolen, M.J., & Brennan, R.L. (2014). *Test equating, scaling, and linking: Methods and practices.* Springer. https://doi.org/10.1007/978-1-4939-0317-7

Kolen, M.J., & Hendrickson, A.B. (2013). Scaling, norming, and equating. In K. F. Geisinger et al. (Eds.), *In APA handbook of testing and assessment in psychology, Vol. 1: Test theory and testing and assessment in industrial and organizational psychology* (pp. 201-222). American Psychological Association.

Kumlu, G. (2019). *Test ve alt testlerde eşitlemenin farklı koşullar açısından incelenmesi* [*An investigation of test and sub-tests equating in terms of different conditions]*. [Doctoral Dissertation, Hacettepe University]. Higher Education Institution National Thesis Center.

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

Lee, W.C., & Brossman, B.G. (2012). Observed score equating for mixed-format tests using a simple-structure multidimensional IRT framework. In M. J. Kolen & W. C. Lee (Eds.), *Mixed-format tests: Psychometric properties with a primary focus on equating* (Volume 2) (CASMA Monograph No. 2.2.) Center for Advanced Studies in Measurement and Assessment, The University of Iowa. http://www.education.uiowa.edu/casma

Lee, G., & Lee, W.C. (2016). Bi-factor MIRT observed-score equating for mixed-format tests. *Applied Measurement in Education*, *29*(3), 224-241. https://doi.org/10.1080/08957347.2016.1171770

Lee, E., Lee, W., & Brennan, R.L. (2014). *Equating multidimensional tests under a random groups design: A comparison of various equating procedures*. (CASMA Research Report No. 40). Iowa City, IA: Center for Advanced Studies in Measurement and Assessment, The University of Iowa. http://www.education.uiowa.edu/casma

Lee, G., Lee, W., Kolen, M.J., Park, I. –Y., Kim, D.I., & Yang, J.S. (2015). Bi-factor MIRT true-score equating for testlet-based tests. *Journal of Educational Evaluation, 28*, 681-700.

Li, Y.H., & Lissitz, R.W. (2000). An evaluation of the accuracy of multidimensional IRT linking. *Applied Psychological Measurement, 24*(2), 115-138.

Lim, E. (2016). *Subscore equating with the random groups design*. [Doctoral dissertation, Graduate College of The University of Iowa]. ProQuest Dissertations & Theses Global.

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.

Livingston, S.A. (2014). *Equating test scores (without IRT)* (2th ed.). Educational testing service, ETS.

Livingston, S.A., & Kim, S. (2009). The circle-arc method for equating in small samples. *Journal of Educational Measurement,46*(3*),* 330-343. https://doi.org/10.1111/j.1745-3984.2009.00084.x

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

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

Loyd, B.H., & Hoover, H.D. (1980). Vertical Equating Using the Rasch Model. *Journal of Educational Measurement, 17* (3), 179-193.

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

McDonald R.P. (2000). A basis for multidimensional item response theory. *Applied Psychological Measurement, 24*, 99-114.

Mutluer, C. (2021). *Klasik test kuramına ve madde tepki kuramına dayalı test eşitleme yöntemlerinin karşılaştırması: Uluslararası öğrenci değerlendirme programı (PISA) 2012 matematik testi örneği [Comparison of test equating methods based on Classical Test Theory and Item Response Theory: International Student Assessment Program (PISA) 2012 mathematics test case]*. [Doctoral dissertation, Gazi University]. Higher Education Institution National Thesis Center.

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

Öztürk, N., & Anıl, D. (2012). Akademik personel ve lisansüstü eğitimi giriş sınavı puanlarının eşitlenmesi üzerine bir çalışma [A study on equating academic staff and graduate education entrance examination scores]. *Eğitim ve Bilim, 37*(165), 180-193.

Petersen, N.S., Cook, L.L., & Stocking, M.L. (1983). IRT versus conventional equating methods: A comparative study of scale stability. *Journal of Educational Statistics, 8*(2), 137-156.

Peterson, J. (2014). *Multidimensional item response theory observed score equating methods for mixed-format tests.* [Doctoral dissertation, Graduate College of The University of Iowa].University of Iowa's Institutional Repository. https://ir.uiowa.edu/cgi/viewcontent.cgi?article=5418&context=etd

Peterson, J., & Lee, W. (2014). Multidimensional item response theory observed score equating methods for mixed-format tests. In M. J. Kolen & W. Lee (Eds.), *Mixedformat tests: Psychometric properties with a primary focus on equating* (Volume 2) (CASMA Monograph No. 2.3). Center for Advanced Studies in Measurement and Assessment, The University of Iowa. http://www.education.uiowa.edu/casma

Powers, S.J., Hagge, S.L., Wang, W., He, Y., Liu, C., & Kolen, M.J. (2011). Effects of group differences on mixed-format equating, In M. J. Kolen & W. C. Lee (Eds.), *Mixed-format tests: Psychometric properties with a primary focus on equating* (Volume 1, pp. 51-73). Center for Advanced Studies in Measurement and Assessment, The University of Iowa. http://www.education.uiowa.edu/casma.

R Core Team (2019). *R: A language and environment for statistical computing* [Computer software]. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/

Salmaner Doğan, R., & Tan, Ş. (2022). Madde tepki kuramında eşitleme hatalarının belirlenmesinde kullanılan delta ve bootstrap yöntemlerinin çeşitli değişkenlere göre incelenmesi [Investigation of delta and bootstrap methods for calculating error of test equation in IRT in terms of some variables]. *Gazi University Journal of Gazi Educational Faculty (GUJGEF), 42*(2), 1053-1081. https://doi.org/10.17152/gefad.913241

Sansivieri, V., Wiberg, M., & Matteucci, M. (2017). A review of test equating methods with a special focus on IRT-based approaches. *Statistica, 77*(4), 329-352. https://doi.org/10.6092/issn.1973-2201/7066

Sass D.A., & Schmitt T.A. (2010). A comparative investigation of rotation criteria within exploratory factor analysis. *Multivariate Behavioral Research*, 45, 73-103. https://doi.org/10.1080/00273170903504810

Sireci, S.G., Thissen, D., & Wainer, H. (1991). On the reliability of testlet-based tests. *Journal of Educational Measurement*, *28*(3), 237-247.

Skaggs, G. (2005). Accuracy of random groups equating with very small samples. *Journal of Educational Measurement, 42(4)*, 309-330

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

Tabachnick, B.G., & Fidell, L.S. (2013). *Using multivariate statistics (6th ed.)*. Pearson.

Tan, Ş. (2015). Küçük örneklemlerde beta4 ve polynomial loglineer öndüzgünleştirme ve kübik eğri sondüzgünleştirme metotlarının uygunluğu [Accuracy of beta4 presmoothing polynomial loglineer presmoothing and cubic spline postsmoothing methods for small samples]. *Gazi University Journal of Gazi Educational Faculty (GUJGEF), 35*(1), 123-151.

Tanberkan Suna, H. (2018). *Grup değişmezliği özelliğinin farklı eşitleme yöntemlerinde eşitleme fonksiyonları üzerindeki etkisi [The effect of group invariance property on equating functions obtained through various equating methods]*. [Doctoral dissertation, Gazi University]. Higher Education Institution National Thesis Center.

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

Tsai, T.H. (1997, March). *Estimating minimum sample sizes in random groups equating*. Annual Meeting of the National Council on Measurement in Education, Chicago, IL.

Tuerlinckx, F., & De Boeck, P. (2001). The effect of ignoring item interactions on the estimated discrimination parameters in item response theory. *Psychological methods*, *6*(2), 181.

Uğurlu, S. (2020). *Comparison of equating methods for multidimensional tests which contain items with differential item functioning*. [Doctoral dissertation, Hacettepe University]. Higher Education Institution National Thesis Center.

Wainer, H. (1995). Precision and differential item functioning on a testlet-based test: The 1991 Law School Admissions Test as an example. *Applied Measurement in Education*, *8*(2), 157-86.

Wainer, H., & Thissen, D. (1996). How is reliability related to the quality of test scores? What is the effect of local dependence on reliability?. *Educational Measurement: Issues and Practice*, *15*(1), 22-29.

Wainer, H., & Wang, X. (2000). Using a new statistical model for testlets to score TOEFL. *Journal of Educational Measurement*, *37*(3), 203-220.

Wang, S., Zhang, M., & You, S. (2020). A comparison of IRT observed score kernel equating and several equating methods. *Frontiers in psychology*, *11*, 308.

Wang, X. (2012). *Effect of sample size on ırt equating of uni-dimensional tests in common item non-equivalent group design*: *A monte carlo simulation study*. [Doctoral Dissertation, Graduate College of The University of Virginia Tech]. ProQuest Dissertations and Theses Global.

Way, W.D., Ansley, T.N., & Forsyth, R.A. (1988). The comparative effects of compensatory and noncompensatory two-dimensional data on unidimensional IRT estimates. *Applied Psychological Measurement, 12*(3), 239-252. https://doi.org/10.1177/014662168801200303

Woodruff, D.J. (1989). A comparison of three linear equating methods for the common-item nonequivalent-populations design. *Applied psychological measurement, 13*(3), 257- 262.

Yen, W.M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement, 8*, 125–145.

Yen, W.M. (1993). Scaling performance assessments: Strategies for managing local item independence. *Journal of Educational Measurement, 30*, 187–213.

Zhang, Z. (2010). *Comparison of different equating methods and an application to link testlet-based tests*. [Doctoral Dissertation, Graduate College of The University of Chinese]. ProQuest Dissertations and Theses Global.

APPENDIX

Table S1. *Item parameter estimates for unidimensional and SS-MIRT models - PISA 2018 Mathematics Test - Form 1 (Türkiye sample).*

Items	D	a_1	D	a_1	a_2
M1	-.439	.590	-.223	.515	
M2	-.517	.656	-.284	.559	
M3	.451	.284	.123	.275	
M4	.663	.855	.431	.656	
M5	1.249	.481	.541	.440	
M6	1.525	1.102	1.129	.750	
M7	1.032	.946	.709	.696	
M8	2.848	.108	.304	.110	
M9	.636	1.013	.452	.719	
M10	-.087	.792	-.054	.629	
M11	-1.499	.730	-.884	.599	
M12	.531	.665	.294	.562	
M13	1.772	.805	1.111	.635	
M14	1.758	.371	.611	.348	
M15	.344	.496	.153	.451	
M16	1.133	.965	.787		.902
M17	.754	.696	.431		.720
M18	.719	.394	.264		.481
Mean	.715	.663	.327	.529	.701
Std. Deviation	.988	.270	.493	.178	.211

D = Item difficulty index, a_1 = Item discrimination index for the first dimension, a_2 = Item discrimination index for the second dimension

Table S2. *Item residual correlation matrix.*

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1																				
2	.03																			
3	.00	.07																		
4	.05	.01	.01																	
5	.01	.05	.01	.05																
6	.00	.00	.01	.02	.01															
7	.02	.01	.03	.05	.02	.02														
8	.02	.01	.08	.02	.01	.04	.03													
9	.01	.04	.04	.06	.02	.01	.00	.04												
10	.00	.05	.02	.08	.01	.05	.05	.03	.03											
11	.01	.05	.06	.07	.04	.01	.03	.03	.01	.01										
12	.01	.02	.01	.03	.02	.02	.01	.01	.03	.02	.05									
13	.01	.10	.01	.03	.02	.02	.02	.00	.07	.00	.04	.02								
14	.02	.01	.05	.04	.04	.02	.04	.02	.04	.03	.05	.02	.02							
15	.01	.00	.03	.07	.02	.06	.03	.05	.03	.04	.06	.04	.03	.01						
16	.03	.02	.03	.03	.03	.01	.03	.01	.00	.01	.06	.11	.05	.01	.06					
17	.03	.02	.01	.00	.01	.04	.03	.04	.01	.02	.02	.02	.00	.00	.05	.01				
18	.06	.06	.07	.06	.06	.02	.04	.11	.00	.02	.05	.01	.07	.01	.01	.07	.08			
19	.09	.01	.02	.06	.04	.04	.03	.04	.01	.03	.06	.08	.02	.03	.01	.05	.03	.34		
20	.04	.08	.01	.02	.09	.04	.08	.02	.00	.00	.02	.09	.01	.02	.06	.04	.10	.31	.31	

Color coding of residual item correlations: blue indicates negative values, white represents positive values, and green highlights residual item correlations that suggest violations of the local independence assumption.

This table presents the mean standard errors of equating obtained across various sample sizes (250, 1000, and 5000), test lengths (20, 40, and 60 items), and proportions of items loaded onto the second dimension (0%, 15%, 30%, and 50%). Four different equating methods—mean equating, linear equating, unsmoothed equipercentile equating, and smoothed equipercentile equating-were compared.

Table S3. *Mean standard errors of equating obtained from traditional equating methods.*

		Number of items in the test											
N	Item	20				40				60			
	Ratio	Mean	Linear	U-Eq	S-Eq	Mean	Linear	U-Eq	S-Eq	Mean	Linear	U-Eq	S-Eq
250	0%	.0255	.0180	.0342	.0347	.0249	.0303	.0335	.0340	.0253	.0504	.0340	.0354
	15%	.0253	.0167	.0327	.0332	.0247	.0287	.0326	.0331	.0245	.0470	.0318	.0337
	30%	.0241	.0147	.0323	.0326	.0236	.0282	.0313	.032	.0238	.0455	.0307	.0326
	50%	.0238	.0144	.0322	.0324	.0230	.0277	.0303	.0313	.0237	.0452	.0307	.0323
1000	0%	.0132	.0041	.0180	.0181	.0131	.0084	.0175	.0176	.0128	.0119	.0173	.0174
	15%	.0129	.0041	.0173	.0176	.0122	.0077	.0162	.0163	.0120	.0114	.0161	.0163
	30%	.0120	.0039	.0164	.0165	.0115	.0074	.0155	.0157	.0116	.0109	.0157	.0158
	50%	.0120	.0040	.0166	.0164	.0113	.0072	.0154	.0156	.0113	.0107	.0155	.0157
5000	0%	.0058	.0008	.0079	.0079	.0057	.0017	.0078	.0078	.0055	.0025	.0076	.0077
	15%	.0055	.0007	.0073	.0073	.0054	.0015	.0073	.0073	.0053	.0023	.0071	.0071
	30%	.0053	.0007	.0072	.0072	.0052	.0015	.0070	.0070	.0051	.0022	.0069	.0069
	50%	.0053	.0007	.0070	.0071	.0051	.0014	.0070	.0071	.0050	.0022	.0069	.0069

Note. U-Eq = Unsmoothed Equipercentile, S-Eq = Smoothed Equipercentile, Item Ratio = Number of items loaded onto the second dimension

This table presents the mean bias values calculated for different simulation conditions, including variations in sample size, test length, and the proportion of items violating the local independence assumption. Comparisons were made across four traditional equating methods.

Table S4. *Mean bias values obtained from traditional equating methods.*

		Number of items in the test											
	Item	20				40				60			
N	Ratio	Mean	Linear	U-Eq	S-Eq	Mean	Linear	U-Eq	S-Eq	Mean	Linear	U-Eq	S-Eq
250	0%	.0442	.0461	.0439	.0440	-.0375	-.0412	-.0399	-.0398	-.0112	-.0119	-.0116	-.0118
	15%	.0601	.0598	.0550	.0552	-.0266	-.0298	-.0279	-.0294	-.0164	-.0170	-.0157	-.0162
	30%	.0293	.0261	.0243	.0241	-.0102	-.0127	-.0123	-.0133	-.0030	-.0035	-.0026	-.0032
	50%	.0407	.0359	.0326	.0327	.0016	-.0012	-.0029	-.0033	.0019	.0016	.0018	.0014
1000	0%	-.0121	-.0147	-.0147	-.0148	-.0183	-.0181	-.0176	-.0177	-.0008	-.0029	-.0028	-.0028
	15%	-.0190	-.0194	-.0188	-.0191	-.0208	-.0202	-.0188	-.0187	-.0124	-.0124	-.0124	-.0121
	30%	.0097	.0086	.0072	.0072	-.0120	-.0115	-.0103	-.0102	-.0144	-.0149	-.0151	-.0148
	50%	-.0036	-.0032	-.003	-.0032	-.0079	-.0074	-.0070	-.0069	-.0018	-.0027	-.0045	-.0044
5000	0%	-.0339	-.0339	-.0330	-.0331	.0188	.0194	.0188	.0188	.0096	.0097	.0093	.0094
	15%	-.0408	-.0410	-.0386	-.0387	-.0073	-.0078	-.0079	-.0078	.0084	.0102	.0089	.0091
	30%	-.0241	-.0249	-.0229	-.0230	-.0107	-.0113	-.0110	-.0110	.0083	.0102	.0087	.0090
	50%	-.0192	-.0201	-.0195	-.0196	-.0288	-.0298	-.0273	-.0273	.0049	.0062	.0055	.0058

Note. U-Eq = Unsmoothed Equipercentile, S-Eq = Smoothed Equipercentile, Item Ratio = Number of items loaded onto the second dimension

This table presents the mean standard errors of observed and true score equating obtained using the mean-standard deviation, mean-mean, Stocking-Lord, and Haebara scale transformation methods across different sample sizes, test lengths, and levels of multidimensionality.

Table S5. *Standard errors of equating obtained in observed and true score equating using scale transformation methods.*

		Number of items in the test											
	N	Item	20				40				60			
		Ratio	M-M	M-S	H	S.L	M-M	M-S	H	S.L	M-M	M-S	H	S.L
Observed Score Equating	250	0%	.0050	.0143	.0043	.0021	.0033	.0127	.0030	.0014	.0026	.0096	.0026	.0012
		15%	.0076	.0224	.0044	.0024	.0041	.0136	.0033	.0017	.0031	.0116	.0026	.0013
		30%	.0071	.0203	.0042	.0024	.0038	.0130	.0032	.0017	.0027	.0111	.0025	.0012
		50%	.0066	.0174	.0056	.0025	.0038	.0141	.0031	.0016	.003	.0114	.0026	.0013
	1000	0%	.0020	.0076	.0020	.0009	.0013	.0059	.0014	.0006	.0012	.0045	.0013	.0005
		15%	.0032	.0127	.0021	.0011	.0021	.0081	.0016	.0008	.0016	.0052	.0013	.0007
		30%	.0029	.0116	.0020	.0011	.0019	.0067	.0016	.0008	.0014	.005	.0013	.0006
		50%	.0020	.0085	.0018	.0008	.0017	.0061	.0015	.0007	.0015	.0058	.0013	.0007
	5000	0%	.0008	.0034	.0009	.0004	.0006	.0025	.0005	.0003	.0005	.0021	.0005	.0002
		15%	.0012	.0041	.0009	.0005	.0008	.0029	.0007	.0004	.0007	.0025	.0006	.0003
		30%	.0011	.0037	.0009	.0004	.0007	.0029	.0007	.0003	.0006	.0024	.0006	.0003
		50%	.0014	.0039	.0010	.0005	.0008	.0031	.0007	.0003	.0006	.0025	.0005	.0003
True Score Equating	250	0%	.0052	.0152	.0046	.0024	.0034	.0130	.0031	.0016	.0026	.0097	.0026	.0012
		15%	.0081	.0236	.0049	.0030	.0042	.0140	.0034	.0019	.0032	.0118	.0027	.0014
		30%	.0077	.0217	.0048	.0032	.0040	.0133	.0034	.0019	.0028	.0113	.0025	.0013
		50%	.0067	.0186	.0060	.0030	.0039	.0146	.0032	.0017	.0031	.0116	.0026	.0014
	1000	0%	.0020	.0079	.0021	.0010	.0014	.0061	.0014	.0007	.0012	.0046	.0012	.0006
		15%	.0033	.0132	.0022	.0012	.0021	.0083	.0016	.0009	.0016	.0052	.0014	.0007
		30%	.0031	.0121	.0022	.0013	.0020	.0070	.0016	.0009	.0015	.0051	.0013	.0007
		50%	.0021	.0089	.0019	.0010	.0017	.0064	.0015	.0008	.0015	.0059	.0013	.0007
	5000	0%	.0009	.0036	.0009	.0004	.0006	.0025	.0006	.0003	.0005	.0021	.0005	.0002
		15%	.0012	.0045	.0010	.0006	.0009	.0030	.0007	.0004	.0007	.0026	.0006	.0003
		30%	.0011	.0039	.0010	.0005	.0008	.0030	.0007	.0004	.0007	.0024	.0006	.0003
		50%	.0014	.0043	.0011	.0006	.0007	.0033	.0007	.0004	.0006	.0026	.0005	.0003

Note. M-M = Mean-Mean, M-S = Mean-Sigma, H = Haebara, S.L = Stocking and Lord, Item Ratio = Number of items loaded onto the second dimension

This table presents the mean bias values of observed and true score equating obtained using the mean-standard deviation, mean-mean, Stocking-Lord, and Haebara scale transformation methods across different sample sizes, test lengths, and levels of multidimensionality.

Table S6. *Bias values obtained in observed and true score equating using scale transformation methods.*

		Number of items in the test											
	N	Item	20				40				60			
		Ratio	M-M	M-S	H	S.L	M-M	M-S	H	S.L	M-M	M-S	H	S.L
Observed Score Equating	250	0%	-.0001	0,0000	-.0015	-.0003	-.0007	.0004	-.0001	.0002	-.0012	-.0013	.0014	.0005
		15%	-.0101	-.0096	-.0018	.0010	.0032	.0039	-.0034	-.0006	.0024	.0025	-.0009	-.0005
		30%	-.0121	-.0108	.0001	.0019	.0039	.0042	-.0063	-.0014	.0022	.0022	-.0017	-.0008
		50%	-.0028	-.0041	-.0090	.0012	-.0005	.0013	.0007	.0003	-.0025	-.0027	.0045	.0010
	1000	0%	-.0011	.0008	-.0021	.0005	-.0016	-.0018	.0010	.0005	.0016	.0017	-.0014	-.0005
		15%	-.0011	-.0005	.0009	.0003	.0013	.0005	0,0000	-.0003	.0021	.0029	-.0019	-.0006
		30%	-.0055	-.0040	.0033	.0014	.0014	.0009	-.0017	-.0006	.0023	.0028	-.0031	-.0008
		50%	-.0006	.0003	.0013	.0002	-.0018	-.0019	.0028	.0009	0,0000	.0007	.0029	.0002
	5000	0%	-.0002	-.0005	.0024	.0002	.0009	.0009	-.0015	-.0005	-.0006	-.0005	-.0024	.0001
		15%	.0013	.0015	.0008	-.0002	-.0030	-.0024	.0021	.0010	-.0008	-.0007	-.0018	0,0000
		30%	.0034	.0034	-.0027	-.0012	-.0021	-.0017	.0026	.0009	-.0010	-.0007	-.0017	0,0000
		50%	-.0008	-.0020	.0041	.0008	.0015	.0020	-.0020	-.0008	.0010	.0007	-.0036	-.0004
True Score Equating	250	0%	-.0001	.0004	-.0011	-.0003	-.0007	.0002	.0001	.0002	-.0013	-.0013	.0016	.0005
		15%	-.0096	-.0092	-.0013	.0012	.0032	.0037	-.0033	-.0006	.0024	.0025	-.0008	-.0005
		30%	-.0114	-.0102	.0007	.0023	.0039	.0039	-.0062	-.0013	.0022	.0023	-.0016	-.0008
		50%	-.0029	-.0038	-.0088	.0010	-.0005	.0008	.0007	.0002	-.0025	-.0027	.0046	.0010
	1000	0%	-.0011	.0002	-.0018	.0004	-.0016	-.0017	.0011	.0005	.0016	.0017	-.0013	-.0005
		15%	-.0012	-.0006	.0011	.0002	.0013	.0006	.0001	-.0003	.0021	.0026	-.0018	-.0006
		30%	-.0053	-.0043	.0036	.0013	.0014	.0010	-.0014	-.0006	.0022	.0025	-.0029	-.0008
		50%	-.0005	-.0002	.0015	.0002	-.0019	-.0017	.0029	.0009	-.0001	.0004	.0028	.0001
	5000	0%	-.0003	-.0002	.0028	.0002	.0010	.0008	-.0014	-.0005	-.0006	-.0005	-.0024	.0001
		15%	.0013	.0016	.0011	-.0002	-.0029	-.0027	.0022	.0009	-.0007	-.0007	-.0017	0,0000
		30%	.0034	.0038	-.0024	-.0011	-.0020	-.0021	.0028	.0009	-.0009	-.0007	-.0015	0,0000
		50%	-.0008	-.0017	.0042	.0007	.0014	.0017	-.0020	-.0008	.0010	.0008	-.0035	-.0003

Note. M-M = Mean-Mean, M-S = Mean-Sigma, H = Haebara, S.L = Stocking and Lord, Item Ratio = Number of items loaded onto the second dimension