ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
Research Article; 2022; 13(2); 105-116

Investigation of the Missing Data Imputation Methods on Characteristic Curve Transformation Methods Used in Test Equating[*]

Gülden ÖZDEMİR[**] Burcu ATAR[***]

Abstract

In this research, the aim was to evaluate the effect of zero imputation and multiple imputation missing data handling methods on item response theory (IRT) based test equating methods under different conditions. Data in this study was obtained from the administration of the TIMSS 2019 eighth-grade science test. Data sets were formed by randomly selecting a sample of 1000 students with full data from booklets 7 and 8. By deleting data under a completely random missing data mechanism within the scope of common-item nonequivalent groups (CINEG) design, four different data sets were obtained with the missing data rates of 10% or 20% in the new test or in both tests. The missing da ta problem was solved by using zero imputation and multiple imputation methods from these data sets. In this way, 8 different data sets were formed. Then, scaling transformation was performed by using characteristic curve transformation methods (Haebara, Stocking-Lord). Test equating results were reported in terms of observed scores. The root mean square error (RMSE) was used as the evaluation criterion to determine the error involved in test equating. As a result, it was determined that in the case of 10% missing data in both tests, generally lower RMSE values were obtained. It was observed that the multiple imputation method, one of the methods for handling missing data, was the method that produced RMSE values that were both the lowest and closer to the full data set as a reference value compared to the zero-imputation method. In addition, it was determined that, when compared to the Haebara method, Stocking-Lord method, one of the characteristic curve transformation methods, produced lower RMSE values and these values were closer to the full data set, which was taken as a reference value.

Keywords: Missing data, zero imputation method, multiple imputation method, test equating, characteristic curve transformation methods

Introduction

Exams play an important role in making some critical decisions in the lives of individuals. Selection of personnel for an institution, promotion, change of title, determination of level, selection of students for higher education, etc., are among those exams. Such exams are carried out at the national or international level for various purposes. There are some exams that can be administered multiple times a year (ALES, YDS, YÖKDİL, TOEFL, etc.) or in certain cycles (TIMSS, PISA, PIRLS, etc.). For these exams, alternative test forms consisting of different items are being developed to ensure the safety of the items (Cook & Eignor, 1991). Alternative test forms, which are also called parallel test forms, are very difficult to produce. There may be slight differences between the difficulty levels of the forms (Kolen & Brennan, 2004). It is important for validity that such exams treat all individuals who take different test forms equally and impartially (Kan, 2011). For this reason, in order to directly compare the performances of individuals who answered different items, their scores should be placed on a common scale. With this method, called test equating, different test forms are equated and the scores obtained become comparable (Cook & Eignor, 1991).

[*] This study was produced from the first author's doctoral thesis. Also, this study was presented at the 7th International Congress on Measurement and Evaluation in Education and Psychology (September 1-4, 2021, Ankara/Turkey).

[**] Teacher, Ministry of National Education, Ankara-Turkey, gulden.ozdemir@meb.gov.tr, ORCID ID: 0000-0002-3150-9438

[***] Prof. Dr., Hacettepe University, Faculty of Education, Ankara-Turkey, burcua@hacettepe.edu.tr, ORCID ID: 0000-0003-3527-686X

To cite this article:
Özdemir, G., & Atar, B. (2022). Investigation of the missing data imputation methods on characteristic curve transformation methods used in test equating. *Journal of Measurement and Evaluation in Education and Psychology, 13*(2), 105-116. https://doi.org/10.21031/epod.1029044

Received: 27.11.2021
Accepted: 28.04.2022

Different data collection patterns (random groups design, single group design, and common-item nonequivalent groups design) can be used in test equating studies. In this study, the common-item nonequivalent groups (CINEG) design was used since the data set in the study was collected in accordance with this design. At CINEG, different groups take different forms of tests. These test forms have common items. Common items are used to reveal the equating relationship between the two groups by comparing the performances of each group (Hambleton et al., 1991; Kolen & Brennan, 2004). Common items include structure, item type, content, etc. of the entire test. In this respect, it is recommended to have a smaller version (representative) of the test (Angoff, 1971). In studies conducted in the related literature, it was aimed mainly to determine the test equating method that shows better performance under different conditions (such as sample size, number of items, item threshold parameter difference, item parameter drift, differential item functioning, guessing, mixed-format test, etc.), (Atalay Kabasakal, 2014; Aytekin Kazanç, 2019; Demirus, 2015; Han, 2008; Karagül, 2020; Kilmen, 2010; Mutluer, 2013; Tian, 2011; Uysal, 2014; Wolf, 2013) or it was aimed to compare the performances of test equating methods based on Classical Test Theory (CTT) and Item Response Theory (IRT) (Mutluer, 2021; Skaggs, 2005; Yang, 1997). Test equating methods are based on different theories such as CTT and IRT (Ryan & Brockmann, 2009). However, the research results show that test equating methods based on IRT generally give better results than methods based on CTT, depending on the sample size and the number of items (Hambleton & Jones, 1993; Jabrayilov et al., 2016; Mutluer, 2021; Yang, 1997).

Different test equating methods are used in IRT. Accordingly, this method can be examined under two headings: concurrent calibration method and separate calibration method. In the concurrent calibration method, item parameters are estimated together for both test forms. The estimated parameters are automatically on the same scale. In the separate calibration method, item parameters are estimated parameters on different scales, and linking or a scale transformation is needed. These transformation methods are referred to as moment methods (mean-mean, mean-sigma) and characteristic curve methods (Haebara and Stocking Lord; Kolen & Brennan, 2004). The mean-mean method described by Loyd and Hoover (1980) calculates by using the means of discrimination (*a*) and difficulty (*b*) parameters. Thus, A slope and B constant values are obtained, which help to determine the individual's ability levels in different test forms. Mean-sigma method described by Marco (1977) calculates by using the mean and standard deviation values of the *b* parameter. Thus, the coefficients A slope and B constant are determined. In the characteristic curve transformation methods developed by Haebara (1980) and Stocking and Lord (1983), parameters *a, b* and *c* (chance parameter) are estimated simultaneously. According to the Haebara (1980) approach, the difference between the item characteristic curves is a function that gives the sum of the squares of the differences between the item characteristic curves of each item for respondents at a given ability level. In the function developed by Stocking and Lord (1983), it is the square of the sum of the difference between the item characteristic curves of each item for respondents at a certain ability level. Whichever of these methods is used, IRT equating is performed after the item calibration and scale transformation steps. Test forms can be used interchangeably as a result of test equating, but proof of validity must be submitted for each alternative form used in national or international exams where important decisions about individuals will be made.

Missing data is an essential factor in making critical decisions about individuals, which may pose a question mark about test validity (Hohensinn & Kubinger, 2011). Missing data occurs as a result of not answering some of the items in the exams or leaving them blank. Missing data may cause a narrowing in the data set, as well as weakening the power of the estimations to be made (Rubin, 1987). On the other hand, there are also studies on missing data such as internal consistency, variance analysis parameters, model-data fit and item-data fit, psychometric properties of scales, measurement invariance, and changing item function affect (Akbaş, 2014; Bayhan, 2018; Enders, 2004; Hohensinn & Kubinger, 2011; Işıkoğlu, 2017, Öztemür, 2014; Tamcı, 2018). In addition, standard analysis methods are prepared according to the full data set and cannot be applied to missing data sets (Rubin, 1987).

Missing data can be on three different missing data mechanisms such as missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR), depending on whether the probability of missing data in a variable is related to other variables. MCAR data mechanism is the situation where there is no relationship between the probability of missing data in a variable and the values of this variable and the other variables; that is, it is completely random (Enders, 2010). In this case, the missing data can be negligible provided that the MCAR assumptions are met, but providing the MAR or MNAR assumptions does not provide sufficient evidence for the negligibility of the missing data. Therefore, when missing data is not negligible, it is necessary to use an appropriate method of handling missing data in the analyses to be made regarding the psychometric properties of the tests (Demir, 2013). For this reason, different methods for handling missing data have been developed. These methods were described by Little and Rubin (2002) as deletion methods (listwise deletion, partwise deletion), imputation methods (average imputation, regression imputation, hot/cold deck imputation, etc.), and model-based methods (expectation-maximization, multiple imputation method, Bayesian imputation methods, etc.).

The statistical software used in the researches offers treating as not administered or treating as incorrect as the default method for processing missing data (Ertoprak, 2017). In the method of treating as incorrect, also called the zero imputation method, if the value of 0 is among the values that can be obtained by observation in the data set, then the value of 0 is imputed instead of the missing data (McKnight et al., 2007). In the multiple imputation method, which is one of the newer and probability-based approaches, two or more values are imputed to replace the missing data, reflecting the distribution of possible values (Rubin, 1987). In studies comparing the performances of different methods, the method with the best performance; it has been determined that the rate of missing data varies according to different conditions such as missing data mechanism and sample size (Akbaş, 2014; Allison, 2003; Koçak, 2016; Wu et al., 2015).

Missing data influences the test equating results performed on forms that use different missing data handling methods. In order to equate different test forms with or without error, the data set should be analyzed using the most appropriate missing data handling method. Numerous studies have been encountered on methods of handling missing data or test equating, but it has been observed that studies that deal with both concepts are limited (Ertoprak, 2017; Kim, 2015; Ngudgratoke, 2009; Shin, 2009). When these studies are examined, it has been found that these studies are limited to the 3-parameter logistic model (3PLM), one of the characteristic curve transformation methods based on the Item Response Theory (IRT), and Stocking-Lord (SL) and root mean square error (RMSE) and equating bias (BIAS) values. Most of these studies used simulated data. Therefore, in this study, it is aimed to examine the impacts of the missing data imputation methods on characteristic curve transformation methods used in test equating under different conditions on the real data set.

In this regard, the research questions addressed in this study are:

1. When the test forms obtained by applying the zero imputation method are equated according to the characteristic curve transformation methods, how do the RMSE values change according to the location of missing data in the test forms (both tests, the new test) and the missing data rate (10%, 20%)?
2. When the test forms obtained by applying the multiple imputation method are equated according to the characteristic curve transformation methods, how do the RMSE values change according to the location of the missing data in the test forms (both tests, the new test) and the missing data rate (10%, 20%)?

Method

In this study, it was aimed to determine the effect of missing data coping methods on test equating methods under different conditions. For this purpose, data sets in which the method of dealing with missing data was applied according to the determined conditions were produced and it was planned to find the method that gave the least error. In the research, equating methods are compared with real data sets under different conditions in a controlled manner. The research that contributes to the theory is basic research in this respect (Karasar, 2009).

Data Set

In this study, the Trends in International Mathematics and Science Study (TIMSS) 2019 data were employed. For the study, the top ten countries (Singapore, Taiwan, South Korea, Russia, Finland, Lithuania, Hungary, United States, Sweden, and Portugal) that administer computer-based applications (eTIMSS) at the eighth-grade level in the TIMSS 2019 science achievement test were selected. Then, all the booklets were examined, and the booklets numbered 7 and 8, which had the highest number of items, were scored dichotomously, and the number of common items, and the number of respondents, were used. After the student answers containing missing data were removed, 2249 student data were obtained for booklet 7, and 2277 student data were obtained for booklet 8. The data set was formed by randomly selecting a sample of 1000 people from these booklets. The sample size of 1000 was selected as a generous number that would provide accurate results and a good baseline for comparison (Swaminathan & Gifford, 1983).

Data Collection Instruments

The data used in the study were obtained from the database (https://timss2019.org/international-database/) published by the International Association for the Evaluation of Educational Assessment (IEA). Half of the countries participating in TIMSS 2019 used the eTIMSS application for the first time. There were 14 student handbooks consisting of eighth-grade math and science items and common items to make connections between the booklets (Mullis et al., 2020). For each of the booklets numbered 7 and 8 used in this study, a total of 25 dichotomously scored science items were selected, 13 of which were common and 12 were non-common. Since the students who responded to the test forms in question were different, the equating pattern of the study was determined as the common test design in the unequal groups. According to Angoff (1971), in equating studies to be carried out in unequal groups, equating errors are to be minimized when the number of common items is equal to at least 20% of the total number of items.

Data Analysis

Since this research is based on IRT, the basic assumptions were tested first. Eigenvalues were calculated for Booklet 7 and Booklet 8 to test the unidimensionality assumption. In both booklets, it was determined that the eigenvalue of the first factor (6.38, 6.20, respectively) was more than three times the eigenvalue of the second factor (1.56, 1.83, respectively). This is an indication that the measured structure is one-dimensional (Büyüköztürk, 2011). Yen’s (1984) Q3 statistic was used to test the local independence assumption. It was determined that the Q3 values calculated for both booklets did not exceed .20. The fact that the Q3 value was calculated based on the correlation between residual values not exceeding .20 provides evidence for local independence (Zenisky et al., 2001). As a result of the preliminary analysis, it was seen that the assumptions of unidimensionality and local independence were supported. Another assumption of IRT is model-data fit. In order to perform a test equating based on IRT between the booklets belonging to the data sets, the model-data fit condition was checked. The purpose of evaluating this fit was to determine how well an IRT model fits the data (Hambleton & Swaminathan, 1985; DeMars, 2010). -2loglikelihood values and chi-square (X^2) statistics were used to determine which IRT model was compatible with the data.

Table 1
Determination of Model Data Fit

	Booklet 7		Booklet 8	
Model	2PLM	3PLM	2PLM	3PLM
-2loglikelihood	28595.60	28560.98	27775.54	27749.93
Number of parameters	74	99	74	99
Difference		34.62		25.61

According to Table 1, 2PLM (two-parameter logistic model) was preferred because the difference in the likelihood obtained from 2PLM and 3PLM (three-parameter logistic model) for both forms was not statistically significant ($p < .01$). For this reason, 2PLM was preferred for parameter estimation in the research.

After meeting the assumptions, data sets were created from the booklets according to the different conditions in the research. From the data sets, data were deleted under the missing completely at random (MCAR) data mechanism via R. Little's MCAR has been tested to see if the missing data in the data sets is completely random. According to the Little's MCAR test result, it was determined that the missing data was MCAR ($p > .05$). Four different data sets were obtained in the new test or in both tests of 10% or 20% missing data. In this study, booklet 7 was determined as "new test (NT)", booklet 8 as "old test (OT)", booklet 7 and booklet 8 as "both tests (BT)". Using the zero imputation and multiple imputation methods from these data sets, 8 different data sets were created to solve the missing data problem. Detailed information about the data sets formed within the scope of the research is given in Table 2.

Table 2

Data Sets Formed Within the Scope of the Research

Sample size	Missing data rate	Missing data location	Techniques of handling missing	
			Zero imputation	Multiple imputation
1000	10%	New test	DS1*	DS2
		Both test	DS3	DS4
	20%	New test	DS5	DS6
		Both test	DS7	DS8

* DS: Data Set

For item parameter estimation, the Expected A Posteriori (EAP) method (Embretson & Reise, 2000), which uses prior distribution information, was utilized. Analyses were performed with the "mirt" package (Chalmers et al., 2021) in the R software. Since the predicted item and ability parameters are in different scales, they should be placed on a common scale; that is, scale transformation should be performed (Kim & Hanson, 2000). In the research, scale transformation was performed by using characteristic curve transformation methods (Haebara *H*, Stocking-Lord *SL*), which is one of the test equating methods based on IRT. The scores obtained from the new form were equal to the scores obtained from the old form. Analyses were performed with the "equateIRT" package (Battauz, 2021) in R and test scores were reported in terms of observed scores. RMSE was used as the evaluation criterion to determine the error involved in test equating. The RMSE index provides a statistic based on the difference between the actual ability level and the predicted ability level. Equation 1 used to calculate the RMSE coefficient is given below. While writing the equation, Harris and Crouse (1993) and Keller and Keller (2011) were utilized.

$$\text{RMSE} = \sqrt{\frac{1}{f}(\sum_{i=1}^{f}(\hat{\theta}_i - \theta_i)^2}, \quad (1)$$

where $\hat{\theta}_i$, predicted skill level; θ_i, actual skill level; f, frequency.

Results

When the test forms obtained by using missing data imputation methods are equated according to test equating methods, the location of missing data in the test forms and RMSE values according to missing data rates are reported in Table 3 and Figure 1, respectively.

Table 3

RMSE Values in Test Equating with Zero Imputation and Multiple Imputation Methods

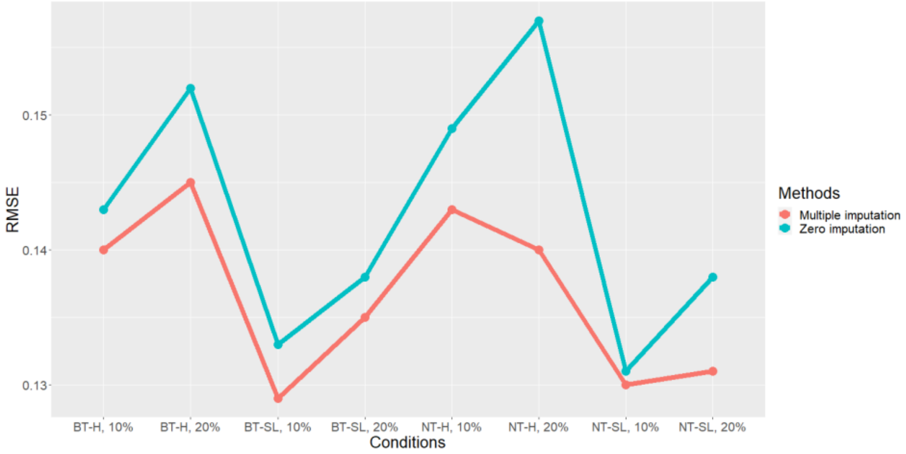
Missing data rate	Missing data imputation methods	Missing data location	Test equating method	Observed score RMSE
Full data set*			Haebara	0.141
			Stocking-Lord	0.130
10%	Zero imputation	New test	Haebara	0.149
			Stocking-Lord	0.131
		Both test	Haebara	0.143
			Stocking-Lord	0.133
	Multiple imputation	New test	Haebara	0.143
			Stocking-Lord	0.130
		Both test	Haebara	0.140
			Stocking-Lord	0.129
20%	Zero imputation	New test	Haebara	0.157
			Stocking-Lord	0.138
		Both test	Haebara	0.152
			Stocking-Lord	0.138
	Multiple imputation	New test	Haebara	0.140
			Stocking-Lord	0.131
		Both test	Haebara	0.145
			Stocking-Lord	0.135

* Taken as a reference value.

According to Table 3, the RMSE value was determined as 0.141 when the test forms that did not contain missing data at the beginning and had full data were equated according to the Haebara method, and the RMSE value was observed as 0.130 when they were equated according to the Stocking Lord method. These values are considered reference values.

Figure 1

RMSE Values in Test Equating with Zero Imputation and Multiple Imputation Methods



According to Figure 1, it is seen that the RMSE values are lower when the test forms obtained by using the multiple imputation method compared to the zero imputation method are equating under all conditions.

Change According to Zero Imputation and Characteristic Curve Transformation Methods

According to Table 3, if 10% missing data was only included in the new test, the RMSE value was determined as 0.149 for test forms with full data obtained by applying the zero imputation method to these missing data when equated according to the Haebara method, and 0.131 when equated according to the Stocking Lord method. In the case of 10% missing data that was included in both tests, the

RMSE value was observed as 0.143 for test forms with full data obtained by applying the zero imputation method to the missing data when equated according to the Haebara method, and 0.133 when equated according to the Stocking Lord method.

In the case of 20% missing data that was only included in the new test, the RMSE value was determined as 0.157 for test forms with full data obtained by applying the zero-imputation method to the missing data when equated according to the Haebara method, and 0.138 when equated according to the Stocking Lord method. In the case of 20% missing data that was included in both tests for test forms with full data obtained by applying the zero imputation method to the missing data, the RMSE value was found as 0.152 when equated according to the Haebara method, and 0.138 when equated according to the Stocking Lord method.

According to Figure 1, when the missing data rate increased, the performance of the zero imputation method decreased and it produced higher RMSE values. In general, it is seen that the missing data is found at the rate of 10% in the new test and the lowest RMSE value is obtained for the condition where the test equating is made according to the Stocking Lord method. In addition, it was determined that the missing data was found at the rate of 20% in the new test and the highest RMSE value was produced for the condition in which the test equating was made according to the Haebara method.

Change According to Multiple Imputation and Characteristic Curve Transformation Methods

According to Table 3, if 10% missing data was only included in the new test for test forms with full data obtained by applying the multiple imputation method to the missing data, the RMSE value was determined as 0.143 when equated according to the Haebara method, and 0.130 when equated according to the Stocking Lord method. In the case of 10% missing data that was included in both tests, the RMSE value was observed as 0.140 for test forms with full data obtained by applying the multiple imputation method to the missing data when equated according to the Haebara method, and 0.129 when equated according to the Stocking Lord method.

In the case of 20% missing data that was only included in the new test for test forms with full data obtained by applying the multiple imputation method to the missing data, the RMSE value was determined as 0.140 when equated according to the Haebara method, and 0.131 when equated according to the Stocking Lord method. When 20% missing data were included in both tests, the RMSE value was found as 0.145 for test forms with full data obtained by applying the multiple imputation method to the missing data when equated according to the Haebara method, and 0.135 when equated according to the Stocking Lord method.

According to Figure 1, when the missing data rate increases, the performance of the multiple imputation method decreases, and it produces higher RMSE values. It is seen that the missing data was found at the rate of 10% in both tests, and the lowest RMSE value was obtained for the condition where the test equating was made according to the Stocking Lord method. In addition, it was determined that the missing data was found at the rate of 20% in both tests, and the highest RMSE value was produced for the condition where the test equalization was made according to the Haebara method.

Discussion and Conclusion

In this research, when the test forms obtained by zero imputation and multiple imputation methods are equated according to characteristic curve transformation methods, one of the test equating methods based on IRT how the RMSE value changes according to different conditions (the rate of missing data, the location of missing data in the test forms) has been examined on the real data set. In the light of the findings obtained from the research, the impact of each missing data handling method on the test equating methods under different conditions was examined and discussed.

When the test forms obtained by applying the zero-imputation method were equated according to the characteristic curve transformation methods, the lowest RMSE value was obtained with 10% of missing data in both tests. In addition, it was observed that RMSE values increased when the missing data rate was 20%. When the test forms obtained by applying the multiple imputation method were equated according to the characteristic curve transformation methods, the lowest RMSE value was obtained with 10% of missing data in both tests, and it was obtained with 20% of missing data in the new test. These results indicate that lower RMSE values were generally obtained when missing data were present in both tests (new form, old form). These indicators support the results of the study conducted by Ertoprak (2017), which also covers the conditions in which the missing data was found in both tests, the new test and the joint test. According to the missing data rate condition discussed in the research, it was seen that the equating error generally increases as the amount of missing data increases. This result concurs with studies showing that the data sets handled under different conditions give more reliable results as the missing data rate decreases, RMSE and bias values decrease, and a closer parameter estimation can be made (Bayram, 2020; Finch, 2008; Zhu, 2014).

The multiple imputation method has been determined as the method that produces the lowest RMSE value among the methods of handling missing data discussed in the research. In addition, it was concluded that the multiple imputation method produced RMSE values closer to the full data set, which was considered the reference value in the research. This result coincides with the findings in the studies on the methods of handling missing data in the literature. It was emphasized that under the conditions discussed in the studies, the multiple imputation method came to the fore because it produced fewer error values (Bayram, 2020; Demir, 2013; Koçak, 2016; Zhu, 2014). Additionally, the results of the studies on methods of handling missing data and test equating methods together support this result (Ertoprak, 2017; Kim, 2015; Ngudgratoke, 2009; Shin, 2009).

The Stocking-Lord method, one of the characteristic curve transformation methods among the test equating methods examined in the research, produced both the lowest and the closest RMSE value to the full data set, which was considered as a reference value compared to the Haebara method. This result is consistent with the results of the study conducted by Karkee and Wright (2004), Kilmen (2010), Aksekioğlu (2017) and Mutluer (2021), which found that the Stocking-Lord method outperformed the Haebara method. However, in the study conducted by Lee and Ban (2010) using a random group design, they found that the Haebara method gave better results than the Stocking-Lord method. The reason for this can be explained by the difference in the selected test equating pattern.

Based on these results, in order to make an equation with the data set that is scored dichotomously and contains missing data, before determining the most appropriate techniques of handling the missing data, it is regarded as essential to examine the missing data rate and the location of the missing data in the test forms. As the missing data rate increases, the performance of the methods of dealing with missing data decreases and the RMSE values increase. According to the conditions discussed in the research, the multiple imputation method, one of the methods for dealing with missing data, and Stocking-Lord method, one of the test equating methods, came to the fore as less error-producing methods. However, it should also be noted that there is no single method that can be used in all conditions and gives the best results. This research is limited to a single data set as it has been obtained from the real data set. For this reason, it is recommended to conduct different studies to compare the results by replication. In addition, for further research, it can be suggested that this study should be conducted using different sample sizes, missing data mechanisms, methods of handling missing data, equating design, test equating methods, and/or evaluation criteria.

Declarations

Conflict of Interest: No potential conflict of interest was reported by the authors.

Ethical Approval: Secondary data were used in this study. Therefore, ethical approval is not required.

References

Akbaş, U. (2014). *Farklı örneklem büyüklüklerinde ve kayıp veri örüntülerinde ölçeklerin psikometrik özelliklerinin kayıp veri baş etme teknikleri ile incelenmesi [Investigation of psychometric properties of scales with missing data techniques for different sample sizes and missing data patterns]* (Thesis No. 370326) [Doctoral dissertation, Ankara University]. Council of Higher Education Thesis Center. https://tez.yok.gov.tr/UlusalTezMerkezi/

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]* (Thesis No. 454879) [Master thesis, Akdeniz University]. Council of Higher Education Thesis Center. https://tez.yok.gov.tr/UlusalTezMerkezi/

Allison, P. D. (2003). Missing data techniques for structural equation modeling. *Journal of Abnormal Psychology, 112*(4), 545-557. https://doi.org/10.1037/0021-843X.112.4.545

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

Atalay Kabasakal, K. (2014). *Değişen madde fonksiyonunun test eşitlemeye etkisi [The effect of differential item functioning on test equating]* (Thesis No. 363206) [Doctoral dissertation, Hacettepe University]. Council of Higher Education Thesis Center. https://tez.yok.gov.tr/UlusalTezMerkezi/

Aytekin Kazanç, E. (2019). *Şans başarısının test eşitlemeye etkisinin farklı eşitleme teknikleriyle araştırılması [Investigation of the effect of guessing on test equating with different equating methods]* (Thesis No. 584263) [Master thesis, Hacettepe University]. Council of Higher Education Thesis Center. https://tez.yok.gov.tr/UlusalTezMerkezi/

Battauz, M. (2021). *Package 'equateIRT'*. https://cran.r-project.org/web/packages/equateIRT/equateIRT.pdf

Bayhan, A. (2018). *Farklı koşullardaki kayıp veri oranının iç tutarlığa etkisi [The effect of missing data rate on internal consistency within different conditions]* (Thesis No. 531022) [Master thesis, Hacettepe University]. Council of Higher Education Thesis Center. https://tez.yok.gov.tr/UlusalTezMerkezi/

Bayram, İ. (2020). *Kayıp veri ile baş etme yöntemlerinin güvenirlik kestirimleri üzerine etkisi [Comparison of influence of the missing data handling methods on reliability estimation]* (Thesis No. 634087) [Master thesis, Hacettepe University]. Council of Higher Education Thesis Center. https://tez.yok.gov.tr/UlusalTezMerkezi/

Büyüköztürk, Ş. (2011). *Sosyal bilimler için veri analizi el kitabı [Manual of data analysis for social sciences]*. Pegem A Publishing.

Chalmers, P., Pritikin, J., Robitzsch, A., Zoltak, M., Kim, K. H., Falk, C. F., Meade, A., Schneider, L., King, D., Liu, C. W., & Ogreten, O. (2021). *Package 'mirt'*. https://cran.r-project.org/web/packages/mirt/mirt.pdf

Cook, L. L., & Eignor, D. R. (1991). An NCME instructional module on IRT equating methods. *Educational Measurement: Issues and Practice, 10*(3), 37-45. http://www.edmeasurement.net/8225/Cook-1991-equating-methods.pdf

DeMars, C. (2010). *Item response theory: Understanding statistics measurement.* Oxford Press.

Demir, E. (2013). *Kayıp verilerin varlığında iki kategorili puanlanan maddelerden oluşan testlerin psikometrik özelliklerinin incelenmesi [Psychometric properties of tests composed of dichotomous items in the presence of missing data]* (Thesis No. 342477) [Doctoral dissertation, Ankara University]. Council of Higher Education Thesis Center. https://tez.yok.gov.tr/UlusalTezMerkezi/

Demirus, K. B. (2015). *Ortak maddelerin değişen madde fonksiyonu gösterip göstermemesi durumunda test eşitlemeye etkisinin farklı yöntemlerle incelenmesi [The study of the effect of anchor items showing or not showing differantial item functioning to test equating using various methods]* (Thesis No. 399468) [Doctoral dissertation, Hacettepe University]. Council of Higher Education Thesis Center. https://tez.yok.gov.tr/UlusalTezMerkezi/

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

Enders, C. K. (2004). The impact of missing data on sample reliability estimates: Implications for reliability reporting practices. *Educational and Psychological Measurement, 64*(3), 419-436. https://doi.org/10.1177/0013164403261050

Enders, C. K. (2010). *Applied missing data analysis.* The Guildford Press.

Ertoprak, D. G. (2017). *Kayıp verinin test eşitlemeye etkisinin incelenmesi [Investigating the effect of missing data on test equating]* (Thesis No. 470015) [Doctoral dissertation, Hacettepe University]. Council of Higher Education Thesis Center. https://tez.yok.gov.tr/UlusalTezMerkezi/

Finch, H. (2008). Estimation of item response theory parameters in the presence of missing data. *Journal of Educational Measurement, 45*(3), 225-245. https://doi.org/10.1111/j.1745-3984.2008.00062.x

Haebara, T. (1980). Equating logistic ability scales by a weighted least squares method. *Japanese Psychological Research, 22*(3), 144-149. https://doi.org/10.4992/psycholres1954.22.144

Hambleton, R. K., & Jones, RW. (1993). Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: Issues and Practice, 12*(3), 38-47. https://doi.org/10.1111/j.1745-3992.1993.tb00543.x

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory. Principals and applications.* Kluwer Academic Publishers.

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

Han, K. T. (2008). *Impact of item parameter drift on test equating and proficiency estimates* (UMI No. 3325324) [Doctoral dissertation, University of Massachusetts Amherst]. Available from ProQuest Dissertations and Theses Global database.

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

Hohensinn, C., & Kubinger, K.D. (2011). On the impact of missing values on the item fit and the model validness of the rasch model. *Psychological Test and Assessment Modeling, 53*(3), 380-393. https://www.researchgate.net/publication/263655931

Işıkoğlu, M. A. (2017). *Kayıp veri ile baş etme yöntemlerinin ölçme değişmezliğine etkisi açısından karşılaştırılması [Comparison of influence of the missing data handling methods on measurement invariance]* (Thesis No. 484106) [Master thesis, Hacettepe University]. Council of Higher Education Thesis Center. https://tez.yok.gov.tr/UlusalTezMerkezi/

Jabrayilov, R., Emons, W. H. M., & Sijtsma, K. (2016). Comparison of classical test theory and item response theory in individual change assessment. *Applied Psychological Measurement, 40*(8), 559-572. https://doi.org/10.1177/0146621616664046

Kan, A. (2011). Test eşitleme: OKS testlerinin istatistiksel eşitliğinin sınanması [Test equating: Checking statistical equivalance of OKS test edition]. *Education and Science, 36*(160), 38-51. http://eb.ted.org.tr/index.php/EB/article/view/310/258

Karagül, A. E. (2020). *Küçük örneklemelerde çok kategorili puanlanan maddelerden oluşan testlerde klasik test eşitleme yöntemlerinin karşılaştırılması [Comparison of classical test equating methods with polytomously scored tests and small samples]* (Thesis No. 610685) [Master thesis, Ankara University]. Council of Higher Education Thesis Center. https://tez.yok.gov.tr/UlusalTezMerkezi/

Karasar, N. (2009). *Bilimsel araştırma yöntemi [Scientific research method]*. Nobel Publishing.

Karkee, T. B., & Wright, K. R. (2004, April 16). *Evaluation of linking methods for placing three-parameter logistic item parameter estimates onto a one-parameter scale* [Conference presentation]. Annual Meeting of the American Educational Research Association in San Diego, California. https://files.eric.ed.gov/fulltext/ED491663.pdf

Keller, L. A., & Keller, R. R. (2011). The long-term sustainability of different Item Response Theory scaling methods. *Educational and Psychological Measurement, 71*(2), 362-379. https://doi.org/10.1177/0013164410375111

Kilmen, S. (2010). *Madde Tepki Kuramı'na dayalı test eşitleme yöntemlerinden kestirilen eşitleme hatalarının örneklem büyüklüğü ve yetenek dağılımına göre karşılaştırılması [Comparison of equating errors estimated from test equation methods based on Item Response Theory according to the sample size and ability distribution]* (Thesis No. 279926) [Doctoral dissertation, Ankara University]. Council of Higher Education Thesis Center. https://tez.yok.gov.tr/UlusalTezMerkezi/

Kim, J. S., & Hanson B. A. (2000). Test equating under the multiple-choice model. *Applied Psychological Measurement, 26*(3), 255-270. https://doi.org/10.1177/0146621602026003002

Kim, M. S. (2015). *Linking with planned missing data: Concurrent calibration with multiple imputation* [Doctoral dissertation, University of Kansas]. KU Scholar Works. https://kuscholarworks.ku.edu/handle/1808/20985

Koçak, D. (2016). *Kayıp veriyle baş etme yöntemlerinin madde tepki kuramı bir parametreli lojistik modelinde model veri uyumuna ve standart hataya etkisi [The effect of missing data tecniques in one parameter logistic model of item response theory on model fit and standard error]* (Thesis No. 456676) (Doctoral dissertation, Ankara University]. Council of Higher Education Thesis Center. https://tez.yok.gov.tr/UlusalTezMerkezi/

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices* (2nd ed.). Springer.

Lee, W. C., & Ban, J. C. (2010). A comparison of IRT linking procedures. *Applied Measurement in Education, 23*(1), 23-48. https://doi.org/10.1080/08957340903423537

Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data* (2nd ed.). Wiley Publishing.

Loyd, B. H., & Hoover, H. D. (1980). Vertical equating using the Rasch model. *Journal of Educational Measurement, 17*(3), 179-193. https://doi.org/10.1111/j.1745-3984.1980.tb00825.x

Marco, G. L. (1977). Item characteristic curve solutions to three intractable testing problems. *Journal of Educational Measurement, 14*(2), 139-160. https://doi.org/10.1111/j.1745-3984.1977.tb00033.x

McKnight, P. E., McKnight, K. M., Sidani, S., & Figueredo, A. J. (2007). *Missing data: A gentle introduction*. Guilford Press.

Mullis, I. V. S., Martin, M. O., Foy, P., Kelly, D. L., & Fishbein, B. (2020). TIMSS 2019 *international results in mathematics and science*. Boston College, TIMSS & PIRLS International Study Center. https://timssandpirls.bc.edu/timss2019/international-results/

Mutluer, C. (2013). Y*ıl içinde farklı akademik personel ve lisansüstü eğitimi giriş sınavı (ALES) puanlarına ilişkin bir test eşitleme çalışması [A test equating study concerning to ALES (Academic Personnel and Postgraduate Education Entrance Exam) scores obtained at different times in a year]* (Thesis No. 336323) [Master thesis, Abant İzzet Baysal University]. Council of Higher Education Thesis Center. https://tez.yok.gov.tr/UlusalTezMerkezi/

Mutluer, C. (2021). *Klasik Test Kuramı'na ve Madde Tepki Kuramı'na dayalı test eşitleme yöntemlerinin karşılaştırılması: 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]* (Thesis No. 658052) [Doctoral dissertation, Gazi University]. Council of Higher Education Thesis Center. https://tez.yok.gov.tr/UlusalTezMerkezi/

Ngudgratoke, S. (2009). *An investigation of using collateral information to reduce equating biases of the post-stratification equating method* (UMI No. 3381312) [Doctoral dissertation, Michigan State University]. Available from ProQuest Dissertations and Theses Global database.

Öztemür, B. (2014). *Kayıp veri yöntemlerinin farklı değişkenler altında varyans analizi (t- testi, anova) parametreleri üzerine etkisinin incelenmesi [Examining the effect of missing data methods on variance analysis (t-Test, ANOVA) parameters under different variables]* (Thesis No. 357738) [Master thesis, Abant İzzet Baysal University]. Council of Higher Education Thesis Center. https://tez.yok.gov.tr/UlusalTezMerkezi/

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys.* John Wiley & Sons.

Ryan, J., & Brockmann, F. (2009). *A practitioner's introduction to equating with primers on Classical Test Theory and Item Response Theory.* CCSSO.

Shin, S. H. (2009). How to treat omitted responses in Rasch model-based equating. *Practical Assessment Research & Evaluation, 14*(1), 1-8. https://doi.org/10.7275/x9vv-xg85

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

Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement, 7*(2), 201-210. https://doi.org/10.1177/014662168300700208

Swaminathan, J., & Gifford, J. A. (1983). Estimation of parameters in the three-parameter latent trait model. In D. J. Weiss (Ed.), *New horizons in testing* (pp. 13-30). Academic.

Tamcı, P. (2018). *Kayıp veriyle başa çıkma yöntemlerinin değişen madde fonksiyonu üzerindeki etkisinin incelenmesi [Investigation of the impact of techniques of handling missing data on differential item functioning]* (Thesis No. 517260) [Master thesis, Hacettepe University]. Council of Higher Education Thesis Center. https://tez.yok.gov.tr/UlusalTezMerkezi/

Tian, F. (2011). *A comparison of equating/linking using the Stocking-Lord method and concurrent calibration with mixed-format tests in the non-equivalent groups commonitem design under IRT* [Doctoral dissertation, Boston College]. Boston College Libraries. http://hdl.handle.net/2345/2370

Uysal, İ. (2014). *Madde Tepki Kuramı'na dayalı test eşitleme yöntemlerinin karma modeller üzerinde karşılaştırılması [Comparison of irt test equating methods for mixed format tests]* (Thesis No. 370226) [Master thesis, Abant İzzet Baysal University]. Council of Higher Education Thesis Center. https://tez.yok.gov.tr/UlusalTezMerkezi/

Wolf, R. (2013). *Assessing the impact of characteristics of the test, common-items, and examinees on the preservation of equity properties in mixed-format test equating* [Doctoral dissertation, University of Pittsburgh]. D-Scholarship. http://d-scholarship.pitt.edu/id/eprint/20130

Wu, W., Jia, F., & Enders, C. (2015). A comparison of imputation strategies for ordinal missing data on likert scale variables. *Multivariate Behavioral Research, 50*(5), 484-503. https://doi.org/10.1080/00273171.2015.1022644

Yang, W. L. (1997). *The effects of content homogeneity and equating method on the accuracy of common-item test equating* (UMI No. 9839718) [Doctoral dissertation, Michigan State University]. Available from ProQuest Dissertations and Theses Global database.

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*(2), 125-145. https://doi.org/10.1177/014662168400800201

Zenisky, A. L., Hambleton, R. K., & Sireci, S. G. (2001). Effects of local item dependence on the validity of IRT item, test and ability statistics [Monograph]. *MCAT*, 3-30. https://eric.ed.gov/?id=ED462426

Zhu, X. (2014). Comparison of four methods for handing missing data in longitudinal data analysis through a simulation study. *Open Journal of Statistics, 4*(11), 933-944. http://dx.doi.org/10.4236/ojs.2014.411088