www.ijpes.com

International Journal of Psychology and Educational Studies

ISSN: 2148-9378

Effect of Missing Data on Test Equating Methods Under NEAT Design

Semih AŞİRET[1], Seçil ÖMÜR SÜNBÜL[2]

[1] *Ministry of National Education, Mersin, Türkiye* *0000-0002-0577-2603*

[2] *Faculty of Education, Mersin University, Mersin, Türkiye* *0000-0001-9442-1516*

ARTICLE INFO

Article History
Received 28.03.2023
Received in revised form 17.04.2023
Accepted 08.05.2023
Article Type: Research Article

ABSTRACT

In this study, it was aimed to examine the effect of missing data in different patterns and sizes on test equating methods under the NEAT design for different factors. For this purpose, as part of this study, factors such as sample size, average difficulty level difference between the test forms, difference between the ability distribution, missing data rate, and missing data mechanisms were manipulated. The effects of these factors on the equating error of test equating methods (chained-equipercentile equating, Tucker, frequency estimation equating, and Braun-Holland) were investigated. In the study, two separate sets of 10,000 dichotomous data were generated consistent with a 2-parameter logistic model. While generating data, the MCAR and MAR missing data mechanisms were used. All analyses were conducted by R 4.2.2. As a result of the study, it was seen that the RMSE of the equating methods increased significantly as the missing data rate increased. The results indicate that the RMSE of the equating methods with imputed missing data are reduced compared to equating without imputed missing data. Furthermore, the percentage of missing data, along with the difference between ability levels and the average difficulty difference between forms, was found to significantly affect equating errors in the presence of missing data. Although increasing sample size did not have a significant effect on equating error in the presence of missing data, it did lead to more accurate equating when there was no missing data present.

Keywords:
Test equating, Missing data, NEAT design

1. Introduction

While the individual may leave one or more items incomplete due to not knowing the answer, not being able to access some items within the specified time, forgetting to respond to the item, or refraining from answering, incomplete data may also arise due to reasons not arising from the individual, such as the technical inadequacy of the data collection tool and carelessness in data entry. Such data are referred to as missing data. The problem of missing data is frequently encountered in the measurement of both cognitive and affective behaviors in social sciences. The most important issue in missing data is the rate of missing data and pattern of missing data. As the amount of missing data increases, statistical bias increases (Enders, 2022; Kang, 2013; Little & Rubin, 1987; Schafer, 1997). Rubin (1976) identified three different mechanisms for the appearance of missing data. The first mechanism occurs when missing data are completely random. In missing completely at random (MCAR) mechanisms, the probability of missing data is independent of all observed variables and the variable in which the missing data accur (Schafer, 1997; Little & Rubin, 1987). In missing at random (MAR) mechanisms, the probability of missing data is independent of the variable in which it occurred but is dependent on other observed variable(s) in the research. In missing not at random (MNAR) mechanisms, the probability of missing data depends on the value of the missing variable (Little & Rubin, 1987).

[1]Corresponding author's address: Ministry of National Education, Mersin /Türkiye
e-mail: semihasiret@gmail.com

Citation: Aşiret, S. & Ömür Sünbül,S. (2023). Effect of missing data on test equating methods under NEAT design. *International Journal of Psychology and Educational Studies*, *10*(3), 702-713. https://dx.doi.org/10.52380/ijpes.2023.10.3.1231

Missing data are an important problem for almost all classical and modern statistical methods used for analysis (Çüm & Gelbal, 2015). Therefore, various missing data imputation techniques have been developed to eliminate missing data. Graham (2009) classified missing data imputation methods as traditional and modern. The multiple imputation (MI) method, which is a modern method, was first proposed by Rubin (1987) and has emerged for many missing data problems (Allison, 2003). With multiple imputation methods, *m* missing datasets are imputed for each missing dataset, and then *m* datasets are obtained. For each dataset, similar analyses are performed to obtain the parameter estimates and standard errors, which are combined according to a certain rule. Finally, a single dataset is obtained (Enders, 2022). The multiple imputation method increases the efficiency of the estimation and reduces bias due to missing data (Olinsky et al., 2003). It is more useful than other modern imputation methods in that it can be applied to all types of models (Allison, 2003). Logistic regression, which is one of the MI techniques, was used as a missing data imputation technique because it is suitable for binary data. (White, Royston, & Wood, 2011). In the first step, logistic regression is applied to observed variables, dependent variables, and the observed predictors. This yields a fitted logistic model for the probability that missing data belong to each category of each binary variable. In the next step, the imputed category is determined by randomly drawing a value from the uniform probability distribution (Enders, 2022).

The comparison of measurements obtained from tests under different conditions by different methods or researchers is the basic prerequisite of all sciences (Dorans & Holland, 2000). This condition is also valid for psychological and educational measurements (Kolen & Brennan, 2014). For security reasons, tests and scales are sometimes administered on different dates or in parallel forms at the same time, which may cause various problems. To overcome these problems, the tests need to be equated statistically.

The collected data and data collection method affect the test equating process. Many different designs are used to collect the data. Data collection designs are divided into two categories: designs using common individuals and designs using common items (Aşiret & Ömür Sünbül, 2016). In this study, the Nonequivalent Anchor Test (NEAT) was used. In the NEAT design, although the groups receiving the test forms do not need to have the same ability distribution, separate tests are applied to two different groups selected from different populations. However, both the tests have common items. These common items can be included in the tests (internal anchors) or used as separate test (external anchors). Forms are equated to each other by means of anchor items in both tests. Anchor items can be thought of as miniatures of the test in terms of content. Additionally, the statistical properties of the anchor items should be the same or similar to the statistical properties of the test. In the NEAT design, different forms of the test are applied to different populations; however, there is a need for a single population in which the equating function is specified. Therefore, the two populations were combined and gathered from a single hypothetical population. Braun and Holland (1982) named this population the synthetic population (Kolen, 1985). Data obtained using the NEAT design are subjected to equating methods for the synthetic population. Common items are used to reveal population differences.

The method to be used should be determined after deciding the design to be used in the test-equating process. When the literature is examined, many different methods have been developed for the test equating process, and these methods basically change within the framework of measurement theory. In this study, the Tucker method (Gulliksen, 1950) and Braun-Holland linear method (Braun & Holland, 1982), which are among the traditional linear equating methods used within the scope of NEAT design, as well as frequency estimation and chained equipercentile equating methods (Angoff, 1971), which are among the traditional equipercentile equating (EE) methods, were used. These methods are briefly described below.

1.1. Tucker Method

The Tucker method is a linear equating method used to equate data obtained using the NEAT design. When the test equating process is carried out with the Tucker method, a new (synthetic) group is formed by weighting the groups that receive the test forms (Kolen, 1985). The Tucker method makes two assumptions for estimating the parameters of the equation. The first assumption is the assumption of linear regression. The regressions of X on anchor test (A) and Y on anchor test (A) are linear and the same for both populations. Another assumption is that the conditional variance of the X and Y scores for any score given on the anchor test is the same for both populations (Kolen & Brennan, 2014; Livingston, 2004).

1.2. Braun-Holland Method

The Braun-Holland equating method is a generalization of the Tucker linear equating method, and the assumptions found in the Tucker linear equating method are accepted in this method (Kolen & Brennan, 2014). The difference in the Tucker method is that it can be used in cases where the assumption of linearity of the graph between the scores obtained from the X and Y forms and the scores obtained from the anchor test form is not accepted. However, the Braun-Holland equating method requires more complex calculations than the Tucker linear equating method (Kolen and Brennan, 2014).

1.3. Chained Equipercentile Equating (CEE) Method

By equating the X form in population P to form A and the form A in population Q to the form Y, the function formed by chaining together two equipercentile equating functions is called the chained EE function. The assumption of this method is that the correlation function from form X to form A and from form A to form Y is the same for both populations (Davier, Holland, & Thayer, 2004).

A shortcoming of this method is that equating the long form, to the short form and then the short form to the long form, can create reliability problems. Another problem is that in the process steps, while the first step is applied only to the first population, the second step is only applied to the second population, so it may be a problem to which population the third step will be applied. To overcome this limitation, Holland and Dorans (2006) assumed that equating would not change all weightings in both populations.

1.4. Frequency Estimation Equating (FEE) Method

The FEE method is used under the EE method based on the NEAT design. For each score given in the anchor test, it is assumed that the conditional distribution of the total score of the X and Y tests is the same for the two synthetic groups. This method is used to estimate the cumulative distributions of the X and Y forms in synthetic populations. The percentile ranks are obtained from the cumulative distributions, and the forms are calculated according to the equipercentile method.

The objective of this research is to examine the effect of missing data in different patterns and sizes on test equating under different conditions (average difficulty difference between forms and ability distribution difference) based on the NEAT design. This study was aimed to address the following research questions:

What is the main effect of

1. the sample size,
2. the difference in average difficulty levels between the test forms
3. the difference in ability level
4. the missing data rate

on the RMSE of NEAT design equating methods through the MCAR and MAR missing data mechanisms?

In the literature, it has been observed that different studies exist in the literature regarding examining the performance of the methods used in test equating under various conditions (Gök & Kelecioğlu, 2014; Puhan, 2010) or studies investigating the effects of missing data (Demir, 2013). Many published studies have examined the effect of missing data on test equating (Ertoprak, 2017; Holland, et al., 2008; Liou & Cheng, 1995; Özdemir & Atar, 2022; Sinharay & Holland, 2008). However, to date, the research has not been able to examine the effect of missing data on NEAT design test equating methods. The validity and accuracy of the measurements are important in studies where serious decisions are made about individuals, different forms are applied for security reasons, or exams administered at different times are evaluated for the same purpose. Therefore, this study is expected to advance the field by examining the effect of missing data on test equating methods.

2. Methodology

This section refers to the type of research, factors manipulated within the scope of this research, and their levels, data generation, and analysis.

2.1. Research Model

This study is fundamental research within the framework of simulation studies which aims to examine the effect of finding missing data with different mechanisms and sizes on equating errors in various conditions of NEAT-based equating methods.

2.2. Manipulated Factors and Their Levels Within the Scope of the Research

One of the important factors affecting the equating error is the sample size. While the standard equating error increases as the sample size increases, the systematic error is not directly affected by sample size (Kolen and Brennan, 2014). Harris (1993) stated that larger sample sizes produce better equating. Kolen and Brennan (2014) suggest that the sample size be 400 per form for linear equating and 1500 per form for equipercentile equating in the NEAT design. Therefore, sample size was manipulated as 1000, 2000, and 4000, representing small, medium, and large samples, respectively, to examine the effect of sample size on equating error in cases where missing data exist and are imputed. The difference in average difficulty levels between test forms has a significant effect on estimating the equating error (Heh, 2007). In various studies (Devdass, 2011; Livingston & Kim, 2011; Parshall et al., 1995), the difference in average difficulty levels to be equated varies between 0.0 and 0.75. In this study, the differences in the average difficulty levels were manipulated as small, medium, and high. Therefore, the average difficulty level difference between the test forms was manipulated as 0.1, 0.3, and 0.5. Uysal and Kilmen (2016) stated that when generating ability distributions, the standard deviation is usually kept constant and the mean is manipulated. In this study, the differences between the ability distributions of the groups were manipulated as low (0.1), medium (0.2), and high (0.3). In the literature related to missing data (Bayhan et al., 2022; Çokluk & Kayri, 2011; Demir, 2013), the missing data rate is commonly determined as 5%, 10%, and 15%. In this study, the missing data rates were determined as 5%, 10%, and 20%.

Within the scope of the research, factors such as sample size (1000, 2000, and 4000), average difficulty level difference between the test forms (0.1, 0.3, 0.5), difference between the ability distribution (0.1, 0.2, 0.3), missing data rate (5%, 10%, 20%), and missing data mechanisms (MCAR and MAR) were manipulated. The effect of these manipulated factors on the RMSE of equating methods based on NEAT design (Tucker, Braun-Holland, CEE, and FEE) was examined. These factors and their levels are listed in Table 1.

Table 1. *Factors and Their Levels in the Study*

Factor	Level	Level Values
Sample Size	3	1000 2000 4000
Average difficulty differences between test forms	3	0,1 0,3 0,5
The difference between ability distribution	3	0,1 0,2 0,3
Missing Data Rate	3	5% 10% 20%
Missing Data Mechanisms	2	Missing Completely at Random (MCAR) Missing at Random (MAR)

2.3. Data Collection Tools and Procedure

In this study, data were generated in accordance with the two-parameter logistic model (2PL) from Item Response Theory models, considering manipulated factors and their levels. According to the NEAT design, two separate population datasets of 10,000 people were generated for forms X and Y, scoring 1-0. The ability parameters were drawn from the normal distribution of $\theta \sim N(0, 1)$. Feinberg and Rubright (2016) stated that if the item difficulty parameters are normal, the discrimination parameters are lognormal, and the lower asymptote parameters are logitnormal, it is easy to derive them from a multivariate normal distribution. Therefore, the parameter b was obtained from $b \sim N(0,1)$ of the standard normal distribution (Bulut & Sünbül, 2017; Feinberg & Rubright, 2016) for form X and for form Y by adding 0.1, 0.3 and 0.5 to the values used in form X. The parameter *b* for the anchor form was generated according to the *b* values in the form Y. Hambleton, Swaminathan, and Rogers (1991) claimed that the appropriate range of item discrimination parameters is 0-2. In this study, the parameter *a* for X and Y forms was obtained from a log-normal distribution (1, 0.05). Since the 2PL model was used, the parameter *c* values are set to 0 for forms X, Y, and A. While both forms X and Y consist of 45 items, the anchor form in both forms consists of 15 items. The ability distribution of the individuals who took the form X was manipulated from the N(0,1) normal distribution, and the ability distribution of the individuals who took the form Y was manipulated by adding 0.1, 0.2, and 0.3 to the average ability values in the first population. In addition, sample sizes were manipulated as 1000, 2000, and 4000, and missing data rates were manipulated as 5%, 10%, and 20%, respectively. For all sample sizes, samples were drawn by performing 100 replications. While generating data, the MCAR and MAR mechanisms were used. In data generation, the R 4.2.2 programming language and the "*mice*" (van Buuren & Groothuis-Oudshoorn, 2011) package in R 4.2.2 were used.

2.4. Data Analysis

In the data analysis, complete datasets and missing datasets were generated primarily. For the MCAR mechanism, missing data were randomly generated at the missing data rates determined within the scope of the study. For the MAR missing data mechanism, individuals were grouped as low, medium, and high according to their ability levels, and random missing data were generated at the determined missing data rates, so that more were at the low ability level and less at the high ability level. The Maximum a posteriori (MAP) method was used to estimate ability of examinees.

The equating phase consists of three different analyses. In the first analysis, samples were drawn from the population without missing data and the forms were equated. In the second analysis, the samples were drawn from the population with missing data, but the forms were equated without assigning missing data. In the final analysis, the samples were drawn from the population with missing data, and the forms were equated after the missing data were assigned using the logistic regression method. In all analyses, the samples were randomly selected, and the X form was equated to the Y form. The EE method was used as a criterion equating method for controlling the results obtained at the end of equating. The "*equate*" (Albano, 2016) package in R 4.2.2 was used to equate the test forms. After the equating process, RMSE for each equating method were calculated, and the equating methods were compared. For all sample sizes, samples were drawn by making 100 replications, and test equating was performed.

3. Findings

The following section presents the findings obtained within the framework of the research questions.

3.1. Findings for the Effect of Sample Size

The main effect graphs of the RMSE of the methods with different sample sizes are presented in Figure 1.

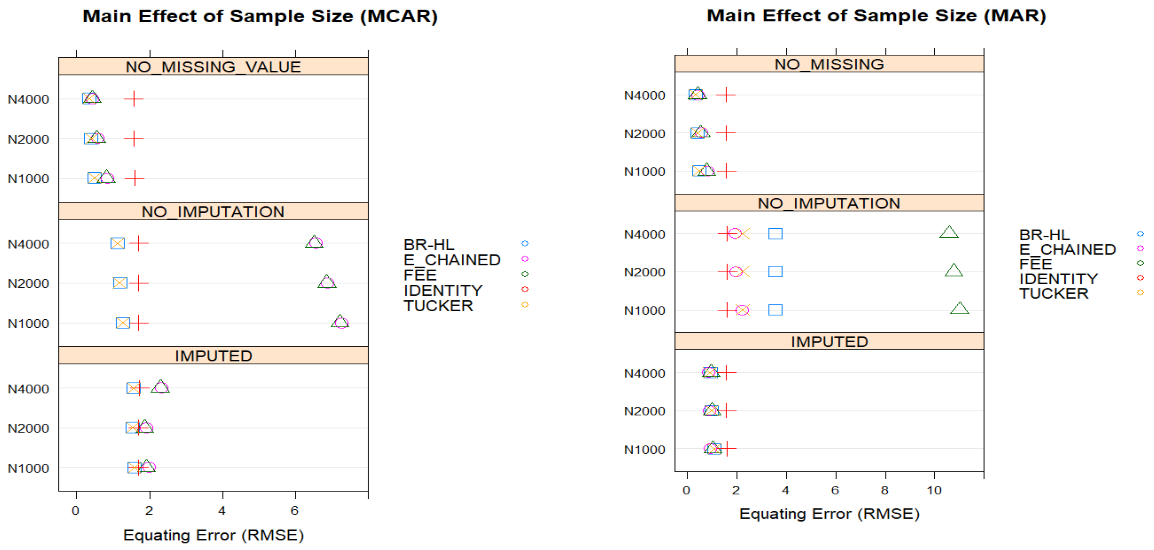

Figure 1. *Main Effects of Sample Size on RMSE for MCAR and MAR Mechanisms*

Figure 1 shows that, when equating with no missing data, the equating errors decrease as the sample size increases for both the MCAR and MAR mechanisms, and all methods have lower equating errors compared to identity equating. It can be stated that the RMSE of the linear equating methods (Braun-Holland and Tucker methods) are lower than those of the nonlinear equating methods (CEE and FEE). This difference is greater with a smaller sample size.

In cases where there are missing data and no missing data imputation is performed, it can be said that the error of the equating methods decreases while the sample size increases. However, while the RMSE of the Tucker and Braun-Holland methods are lower in the MCAR mechanism, they are higher in the MAR mechanism. Although the error value of the FEE method decreases for both missing data mechanisms as the sample size increases, it remains very high. For all sample sizes, the errors in the CEE method are much higher for the MCAR mechanism than for the MAR mechanism.

In cases where missing data imputation is performed, it can be indicated that as the sample size increases, the RMSE decrease with a sample size of 4000 and in the MCAR mechanism, except for the CEE and FEE methods. The RMSE of the Braun-Holland and Tucker methods are lower for the MCAR mechanism than in the identity equating method; in the MAR mechanism, the RMSE of all methods in all sample sizes are lower than those of identity equating.

3.2. Findings for the Effect of the Difference in Average Difficulty Levels between the Test Forms

The main effect of the difference in the average difficulty levels between the test forms on the RMSE of the methods is shown in Figure 2.

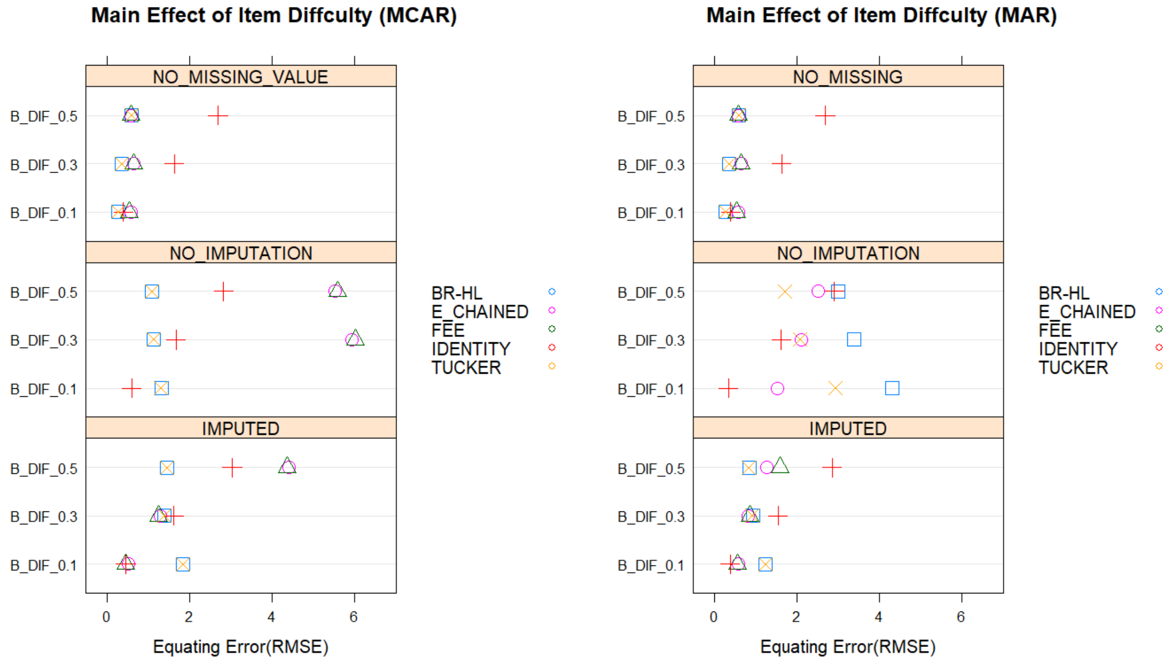

Figure 2. *Main Effects of Average Difficulty Differences Between Test Forms on RMSE for MCAR and MAR Mechanisms*

Figure 2 shows that, in the absence of missing data, the equating errors of all methods increase in both missing data mechanisms, while the RMSE of the identity equating method increase significantly. In addition, while the Braun-Holland and Tucker methods are lower when the average difficulty level between the forms is low, the RMSE of all methods are approximately the same when the average difficulty level difference between the forms is high.

In cases where there are missing data and no missing data imputation is performed, the RMSE of the FEE and CEE methods in the MCAR mechanism are significantly higher. The RMSE of the Braun-Holland and Tucker equating methods do not significantly change. However, the equating errors of the equating methods are significantly high in the MAR missing data mechanism. While the difficulty level between the forms is low, the errors of the equating methods show differentiation. As the average difficulty level between the forms increases, the RMSE of the identity and CEE methods increase, while the RMSE of the other methods decrease. While the equating error of the FEE method does not show significant differentiation as the difficulty level between the forms increases, it is very high compared with other methods.

In cases where missing data imputation is conducted, it can be stated that, for the MCAR mechanism, the RMSE of the identity equating, FEE, and CEE methods increased significantly, whereas those of the Braun-Holland and Tucker methods decreased slightly. In cases where the average difficulty difference between the forms is low, the RMSE of the FEE and CEE methods and identity equating methods are lower. However, the RMSE of the Braun-Holland and Tucker methods are lower as the average difficulty level between forms increases. In the MAR mechanism, the RMSE of the other equating methods are lower than those of the MCAR mechanism, except for the identity equating method. In particular, the errors of the CEE and FEE methods can be said to have decreased significantly.

3.3. Findings for the Effect of the Difference Between Ability Levels

The main effect of the difference between the ability levels on the RMSE of the methods is shown in Figure 3.

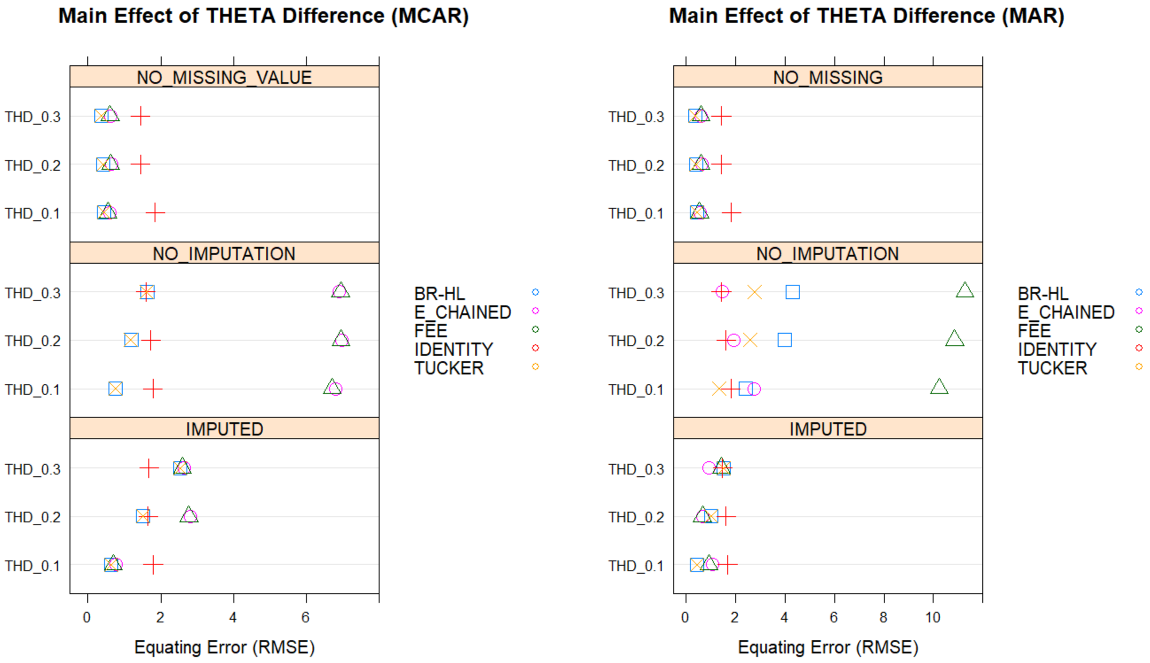

Figure 3. *Main Effects of Difference Between Ability Levels on RMSE for MCAR and MAR Mechanisms*

Figure 3 shows that, in the test with no missing data, the mean equating errors of equating methods for all ability levels are smaller than identity equating. For both missing data mechanisms, the RMSE of the equating methods seem to increase slightly as the difference between the groups increases, and the error values of the Braun-Holland and Tucker methods are lower than those of the CEE and FEE methods.

In cases where there are missing data and no missing data imputation is performed, in the MCAR mechanism, the RMSE of the other methods increased, except for the identity equating method. The RMSE of the CEE and FEE methods are significantly higher. In particular, as the difference in ability levels between the groups increases, the RMSE of the Braun-Holland and Tucker methods and the error values of the identity equating are approximately the same. In the MAR mechanism, on the other hand, the RMSE of all methods are higher than those of the identity equating method. The RMSE of the FEE method is the highest under all conditions. While the RMSE of the Braun-Holland and Tucker methods increase as the ability difference increases, the RMSE of the CEE method decreases.

In cases where missing data imputation is conducted, the RMSE obtained in the MCAR mechanism are higher than those obtained in the MAR mechanism. When the ability level difference is low in the MCAR mechanism, the RMSE of all methods are lower than those with identity equating. As the ability level difference increases, the RMSE of the equating methods also increase. However, the RMSE of the CEE and FEE methods are seen to decrease slightly when the ability difference is 0.3 compared with the difference of 0.2. When the intergroup ability level is 0.3, the RMSE of all methods are higher than those of the identity equating. In the MAR mechanism, on the other hand, the RMSE of the methods are lower than the identity equating under all conditions. In cases where the ability difference is 0.3, the error values of the methods approach the identity equating.

3.4. Findings for the Effect of the Missing Data Rates

The main effect graphs of the RMSE of the methods for the missing data rates are shown in Figure 4. Accordingly, it can be said that, in the absence of missing data, the equating errors of all test equating methods are approximately equal, and they are lower than identity equating.

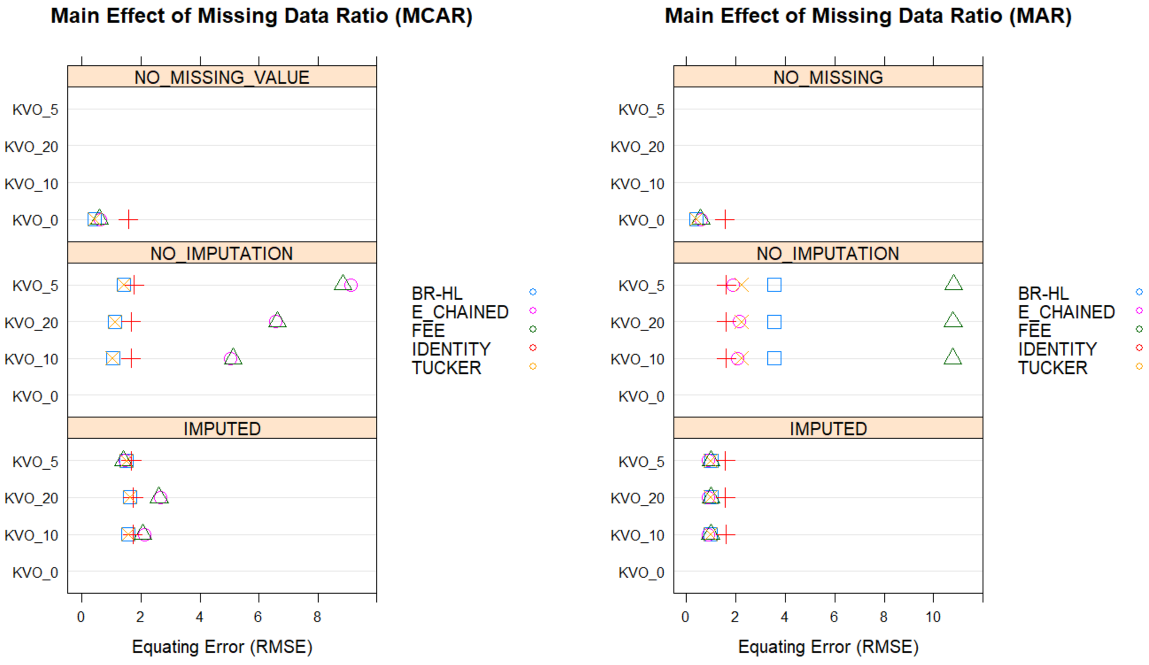

Figure 4. *Main Effects of Missing Data Rate on RMSE for MCAR and MAR Mechanisms*

In the MCAR mechanism, where forms are equated without missing data imputation, the equating errors of the Braun-Holland and Tucker methods are very close to each other and the lowest of all the missing data rates. The CEE and FEE methods are close to each other and have the highest missing data rates. As the missing data rate increases, the equating errors of the Braun-Holland and Tucker equating methods decrease slightly. It is observed that the equating error of the CEE and FEE methods is high when the missing data is low, decreases when the missing data rate is 10%, and increases slightly when the missing data rate is 20%. In the MAR mechanism, on the other hand, as the rate of missing data increases, the equating errors of the methods do not differ significantly. The mean equating error of the identity equating method is the lowest for all missing data rates. This method is followed by the CEE, Tucker, Braun-Holland, and FEE methods, respectively.

After missing data imputation, it can be said that equating errors slightly increase as the missing data rate of all equating methods in the MCAR mechanism increases. When missing data imputation is conducted, the equating errors of the Braun-Holland and Tucker equating methods increase slightly, in line with the missing data rate. The equating errors of these two methods are the lowest for all missing data rates. The increases in the RMSE of the CEE and FEE equating methods are higher than those of the Braun-Holland and Tucker methods. Equating errors in the CEE and FEE methods increase with the rate of missing data increases. While the equating errors of both methods are lower than identity equating at a 5% missing data rate, they are higher than identity equating as the missing data rate increases. In the MAR missing data mechanism, the variation in the missing data rates and the errors of the equating methods do not show a significant difference. The mean equating errors of the methods were significantly reduced when the data imputation was performed. The mean equating error of all methods is lower than identity equating and is very close to each other.

4. Conclusion and Discussion

In the study, the RMSE of test equating methods concerning various variables (missing data mechanism, sample size, difficulty level between forms, ability level differences between the groups, missing data rate) were analyzed and compared. In the study, test equating was performed for the following conditions: there were no missing data, there were missing data, and the missing data was imputed.

One of the factors affecting the RMSE in test equating is sample size. Kolen and Brennan (2014) state that sample size affects the standard error of equating directly, and that error can be reduced by increasing the sample size. In this study, in all circumstances, when sample size increases, the RMSE of identity equating remains stable while the RMSE of other methods decrease.

The difficulty level difference between forms has a significant effect on estimating equating error (Heh, 2007). Kim, von Davier, and Heberman (2006) stated that identity equating would not be appropriate in cases where the difference between forms and score distributions is significant. Similarly, in this study, test equating was performed in situations where there were no missing data or after missing data imputation, and the RMSE of identity equating increased significantly as the difficulty level difference between forms increased. However, in the MCAR mechanism, the RMSE of the CEE and FEE methods after missing data imputation, when the difficulty level difference between forms is 0.5, are higher than identical equating. In this case, using CEE and FEE methods in the MCAR mechanism is not recommended if the difficulty level difference between the forms is high. However, as the difficulty level between forms increases in the MAR mechanism, the RMSE of all methods are lower than identity equating, except for cases where the difficulty level difference between forms is 0.1. Harris and Crouse (1993) and Kolen and Brennan (2014) remarked that in cases where the difficulty levels of test forms are close to each other, not performing equating will cause fewer errors. This is in line with the findings of this study. In this case, in the MAR mechanism, although errors in equating methods increase as the difficulty level between forms increases, they have a lower value of error compared to identity equating. When the RMSE of the equating methods are examined at different difficulty levels after the missing data imputation is performed, the RMSE of the methods are reduced in the MCAR mechanism, while the RMSE of the methods are more affected in the MAR mechanism as the difficulty level between the forms increases. The equating errors of the FEE and CEE methods are more affected by the difficulty level of the forms. It can be said that the RMSE of the Braun-Holland and Tucker methods are less affected by the increase in difficulty level between forms.

Puhan (2010) stated that linear equating methods are affected more by the inability to provide a population invariance assumption. In this study, it is seen that the RMSE of the linear test equating methods (Tucker and Braun-Holland) are higher than the nonlinear equating methods (FEE and CEE) as the difference in ability between the groups increases. This finding is similar to the findings obtained by Puhan (2010).

Enders (2004) stated that the validity and reliability of the measurement procedures performed in cases of missing data are adversely affected. In the study, the RMSE of the test equating methods are lower for both missing mechanisms in the test equating performed without missing data. This is an expected situation. In cases where there is missing data, the FEE and CEE methods are significantly affected by the MCAR mechanism. The RMSE of these methods are quite high. In the MAR mechanism, the RMSE of FEE methods is significantly higher compared to other methods. When the forms are equated after the missing data imputation, in contrast to the MCAR mechanism, the RMSE of all test equating methods in the MAR mechanism are smaller than the RMSE obtained without equating in all sample sizes. Sulis and Porcu (2008) stated that the estimations made in the MCAR mechanism were more effective and more unbiased. Similarly, in the MCAR mechanism in this study, the RMSE of the Braun-Holland and Tucker equating methods are lower in the case of missing data than in the case of missing data in the MAR mechanism. However, the RMSE of the FEE and CEE methods in the MAR mechanism are lower than those in the MCAR mechanism. After missing data imputation in the MAR mechanism, the lowest RMSE belongs to the CEE method. Sinharay and Holland (2010) stated that CEE is the most appropriate method after the item-response-theory observed-score-equating method under the NEAT design in cases of missing data. The findings obtained from this study show parallelism with the findings of Sinharay and Holland (2010).

Özdemir and Atar (2022) compared the RMSE of IRT-based equating methods at 10% and 20% missing data rates, with and without missing data imputation. In the study, they stated that the RMSE of the test equating methods increased as the rate of missing data increased. In this study, it can be said that when the generated forms with different missing data rates are equated, the equating errors increase as the missing data rate increases.

As a result, missing data leads to bias and an increase in error values in test equating, as in other measurement processes. In this study, it can be said that in the MCAR design, the RMSE of linear equating methods increase even when missing data are imputed, but the RMSE of all methods are significantly reduced when nonlinear equating methods and missing data are imputed in the MAR mechanism. Although missing data imputation is carried out, the fact that the researchers obtain data without missing data or with the lowest missing data rate will increase the validity and reliability of the measurements performed. In test equating performed without missing data imputation, the RMSE of nonlinear equating methods (CEE and FEE) are very high in both missing data mechanisms. In this case, it was concluded that nonlinear methods were more affected than linear equating methods.

5. Recommendations

Within the scope of this study, equating was carried out using linear equating methods such as Tucker and Braun-Holland methods and non-linear equating methods using CEE and FEE methods based on NEAT design. Similar studies can be performed for different equating designs and methods. Also, similar studies can be repeated by including the location of the missing data. The MNAR mechanism was not included in the study. Studies can be carried out in accordance with the MNAR mechanism. Ertoprak (2017) suggested the use of multiple imputation techniques based on logistic regression and discriminant functions for dichotomous data sets. In this study, only the multiple imputation method based on logistic regression was used to impute missing data. Similar work can be performed using the multiple imputation method based on the discriminant function.

6. References

Albano, A. D., (2016). equate: An R package for observed-score linking and equating. *Journal of Statistical Software*, 74(8), 1-36. doi:10.18637/jss.v074.i08

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

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

Aşiret, S., & Sünbül, S. Ö. (2016). Investigating test equating methods in small samples through various factors. *Educational Sciences: Theory & Practice*, 16, 647-668.

Bayhan, A., Doğan, N., & Sosyal, S. (2022). The Effect of missing data rate on internal consistency within different conditions. *Mehmet Akif Ersoy Üniversitesi Eğitim Fakültesi Dergisi*. 63, 444-461.

Braun, H. I., & Holland, P. W. (1982). Observed-score test equating: A mathematical analysis of some ETS equating procedures. In P.W. Holland & D. B. Rubin (Eds.), *Test equating* (pp. 9–49). Academic Press.

Bulut, O., & Sünbül, Ö. (2017). Monte Carlo simulation studies in item response theory with the R programming language. *Journal of Measurement and Evaluation in Education and Psychology*, 8(3), 266-287.

Çokluk, Ö., & Kayri, M. (2011). Kayıp değerlere yaklaşık değer atama yöntemlerinin ölçme araçlarının geçerlik ve güvenirliği üzerindeki etkisi. *Kuram ve Uygulamada Eğitim Bilimleri*, 11, 289-309.

Çüm, S.i & Gelbal, S. (2015). Kayıp veriler yerine yaklaşık değer atamada kullanılan farklı yöntemlerin model veri uyumu üzerindeki etkisi. *Mehmet Akif Ersoy Üniversitesi Eğitim Fakültesi Dergisi*, 1(35), 87-111.

Demir, E. (2013). Kayıp verilerin varlığında çoktan seçmeli testlerde madde ve test parametrelerinin kestirilmesi. *Eğitim Bilimleri Araştırma Dergisi*, 3, 47-68

Demir, E. (2013). *Kayıp verilerin varlığında iki kategorili puanlanan maddelerden oluşan testlerin psikometrik özelliklerinin incelenmesi* [Doctoral dissertation]. University of Ankara.

Devdass, S. (2011). *Conditions affecting the accuracy of classical equating methods for small samples under the NEAT design: A simulation study* [Doctoral dissertation]. The University of North Carolina.

Dorans, N. J., & Holland, P. W. (2000). Population invariance and the equatability of tests: Basic theory and the linear case. *Journal of Educational Measurement*, 37, 281–306.

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.

Enders, C. K. (2022). *Applied missing data analysis*. The Guildford Press.

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

Feinberg, R. A., & Rubright, J. D. (2016). Conducting simulation studies in psychometrics. Educational Measurement: Issues and Practice, 35(2), 36-49. https://doi.org/10.1111/emip.12111

Gök, B., & Keleceioğlu, H. (2014). Comparison of IRT equating methods using the common-ıtem nonequivalent groups design. *Mersin University Journal of the Faculty of Education*, *10*(1), pp. 120-136

Gulliksen, H. (1950*). Theory of mental tests*. Wiley.

Hambleton, R.K., Swaminathon, H., Rogers, H. J. (1991). *Fundamentals of ıtem response theory*. Sage.

Harris, D. J. (1993). *Practical issues in equating*. Paper presented at the annual meeting of the American Educational Research Association, Atlanta.

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

Heh, V. K. (2007). *Equating accuracy using small samples in the random groups design* [Doctoral dissertation]. The College of Education of Ohio University.

Heh, V.K., (2007). *Equating accuracy using small samples in the random groups design* [Doctoral dissertation]. The College of Education of Ohio University.

Holland, P. W., & Dorans, N. J. (2006). Linking and equating. In R. L. Brennan (Ed.), *Educational measurement* (pp. 187-220). Praeger Publishers.

Holland, P. W., Sinharay, S., von Davier, A. A., & Han, N. (2008). An approach to evaluating the missing data assumptions of the chain and post-stratification equating methods for the NEAT design. *Journal of Educational Measurement*, 45(1), 17-43.

Kang H. (2013). The prevention and handling of the missing data. *Korean J Anesthesiol*. 64(5), 402-6. doi: 10.4097/kjae.2013.64.5.402.

Kim, S., von Davier, A. A., & Haberman, S. (2006). An alternative to equating with small samples in the non-equivalent groups anchor tests design. *ETS Research Report Series, 2*, 1-40.

Kolen, M. J. (1985). Standard errors of Tucker equating. *Applied Psychological Measurement*, *9*(2), 209-223.

Kolen, M. J.,& Brennan R. L. (2014). *Test equating, scaling, and linking: Method and practic.*, Springer-Verlag.

Liou, M., & Cheng, P. E. (1995). Equipercentile equating via data-imputation techniques. *Psychometrika*, 60(1), 119-136.

Little, R. J. A., & Rubin, D. B. (1987). *Statistical analysis with missing data*. Wiley.

Livingston, S. A. (2004). *Equating test scores (without IRT)*. ETS.

Livingston, S. A.,& Kim, S. (2010), Random-Groups equating with samples of 50 to 400 test takers. *Journal of Educational Measurement*, 47, 175–185.

Özdemir, G., & Atar, B. (2022). Investigation of the missing data ımputation methods on characteristic curve transformation methods used in test equating. *Journal of Measurement and Evaluation in Education and Psychology*, 13-2, 105-116

Parshall, C. G., Houghton, P. D. B., & Kromrey, J. D. (1995). Equating error and statistical bias in small sample linear equating. *Journal of Educational Measurement*, 32(1), 37-54.

Puhan, G. (2010). A comparison of chained linear and poststratification linear equating under different testing conditions. *Journal of Educational Measurement*, 47(1), 54-75.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63(3), 581-592.

Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. Chapman & Hall/CRC.

Sinharay, S., & Holland, P. W. (2008).*The missing data assumptions of the nonequivalent groups with Anchor Test (NEAT) design and their ımplications for test equating*. ETS Research Rep. No. RR-09-16. ETS.

Sinharay, S., & Holland, P. W. (2010). The missing data assumptions of the neat design and their ımplications for test equating. *Psychometrika*, 75- 2, 309–327. https://doi.org/10.1007/S11336-010-9156-6

Sulis, I., & Porcu, M. (2008). *Assessing the effectiveness of a stochastic regression imputation method for ordered categorical data*. Center for North South Economic Research, April 2008.

Uysal, İ. & Kilmen, S. (2016). Comparison of item response theory test equating methods for mixed format tests. *International Online Journal of Educational Sciences*, 8(2), 1-11.

van Buuren S., & Groothuis-Oudshoorn, K. (2011). Mice: multivariate ımputation by chained equations in R. *Journal of Statistical Software*, 45(3), 1-67. https://doi.org/10.18637/jss.v045.i03

von Davier, A. A., Holland, P. W., & Thayer, D. T. (2004). *The kernel method of test equating*. Springer.

White, I. R., Royston, P., & Wood, A. M. (2011). Multiple imputation using chained equations: issues and guidance for practice. *Statistics in Medicine*, 30(4), 377-399.