International Journal of Assessment Tools in Education

2024, Vol. 11, No. 3, 445–462

https://doi.org/10.21449/ijate.1417166

journal homepage: https://dergipark.org.tr/en/pub/ijate *Research Article*

An investigation into the effect of different missing data imputation methods on IRT-based differential item functioning

Fatma Ünal[1*], Hakan Koğar[1]

[1]Akdeniz University, Faculty of Education, Department of Educational Sciences, Antalya, Türkiye

ARTICLE HISTORY

Received: Jan. 12, 2024

Accepted: Apr. 28, 2024

Keywords:

Differential item functioning,

Missing data,

Item response theory,

Raju's area measurement,

Likelihood ratio.

Abstract: The purpose of this study is to examine the effect of missing data imputation methods, namely regression imputation (RI), multiple imputation (MI) and k-nearest neighbor (kNN) on differential item functioning (DIF). In this regard, the datasets used in the research were created by deleting some of the data via the missing completely at random mechanism from the complete datasets obtained from 600 students in Türkiye, the United Kingdom, the USA, New Zealand and Australia, who answered booklets 14 and 15 from the PISA 2018 science literacy test. Data imputation was applied to the datasets through missing data using RI, MI and kNN methods and DIF analysis was performed on all datasets in terms of language and gender variables via Lord's χ^2 method, Raju's area measurement method and item response theory likelihood ratio method. DIF results from the complete datasets were taken as a reference and they were compared with the results from other datasets. As a result of the research, values close to 10% of accurate imputation were achieved in the RI method depending on language and gen-der variables. In MI and kNN methods, results closest to the complete datasets were obtained at a rate of 5% depending on the language variable. In the MI method, inaccurate results were obtained in all proportions in terms of the gender variable. For the gender variable, the kNN method gave accurate results at rates of 5% and 10%.

1. INTRODUCTION

Tests developed for the purpose of detecting cognitive or affective characteristics of individuals such as intelligence, achievement, and attitude can be used in many educational studies. According to the scores obtained from the tests used in the field of education, it is possible to examine how much individuals have the characteristics planned to be measured and evaluations can be made based on the results obtained, and important decisions can be raised about individuals (Uyar, 2015; Yılmaz, 2021).

International monitoring studies in education, such as the Program for International Student Assessment (PISA), make it possible for countries to compare their educational status with other countries (MEB, 2019). Thanks to these studies, countries evaluate their education systems and create appropriate policies. PISA is a study conducted in three-year cycles, aimed at evaluating the ability of students aged 15 to reflect the knowledge and skills they have

*CONTACT: Fatma ÜNAL ✉ fatmaunal.452@gmail.com Akdeniz University, Faculty of Education, Department of Educational Sciences, Antalya, Türkiye

e-ISSN: 2148-7456

acquired in daily life by measuring their science literacy, mathematics literacy and reading skills (MEB, 2019).

Science literacy assesses individuals' ability to engage with science-related topics and scientific phenomena. Individuals who have acquired science literacy should have the ability to explain events in a scientific way, design and evaluate scientific work, stand willing to demonstrate their ability to interpret data and evidence scientifically (OECD, 2019).

There are 3 types of information in science literacy: content, method, and epistemic information. PISA focuses on the capacity of 15-year-old students to reveal these types of information in an appropriate way in personal, local, national, and global situations (OECD, 2019). As a result of the PISA application, the knowledge and skill levels of students in a country can be compared with students in other participating countries. At the same time, standards are established to raise the education levels of countries, and the strengths and weaknesses of education systems can be identified (Taş et al., 2016). Based on this, it can be said that some important inferences can be made about education thanks to studies in education such as PISA. For this reason, to make correct inferences, first, accurate results should be obtained from the studies. Among the reasons for making inaccurate comments and corrections on the research results are the decrease in the validity of the research results and the negative impact on validity. Validity is one of the most important features expected in measurement tools and DIF is one of the factors that cause a decrease in validity (Sırgacı & Çakan, 2020).

In the tests applied in the field of education, individuals at the same ability level are expected to get the same scores from the test items. When individuals in different groups at the same ability level score differently on test items, this indicates that the items are biased towards one group. To determine this bias, differential item functioning analyses are performed on the dataset (Atar et al., 2021, p. 419). DIF analyses assume that the same characteristics of individuals in different subgroups are measured in a test. The goal here is to distinguish between real differences between groups and measurement bias (Kalaycıoğlu & Kelecioğlu, 2011). In order to perform DIF analyses, subgroups are first determined in terms of the variables such as language, gender, and race. The responses to the test items should not differ according to these predetermined subgroups but should differ according to the ability levels of individuals. One of the subgroups is selected as the focus group and the other as the reference group. The responses of the individuals to the test items are compared in the focus group and the reference group. If the probability of answering an item correctly differs from one subgroup to another, it is stated that there is DIF in that item (Dogan et al., 2005). There are some situations that cause DIF in an item. These situations include socio-economic level, comprehensibility of the item, curriculum, poor translation, item writing, the relationship between the content and language of the item and culture, the meaning inferred from the item, and differences in sentence structure (Van de Vijver & Tanzer, 1997). DIF can be analyzed with methods based on item response theory and classical test theory.

Item Response Theory (IRT) consists of a mathematical model indicating the relationship between an individual's observable performance on a test and the latent traits or abilities that are thought to underlie this performance (Hambleton & Swaminathan, 2013, p. 9). With this theory, it is stated that under the assumptions of unidimensionality, local independence, and model-data fit, the estimation of ability parameters can be performed independently of the properties of the items and the estimation of item parameters can be obtained independently of the sample of the study (Gültekin & Demirtaşlı, 2020). In item response theory, the qualifications of the individuals in the study are first determined. Then, scores are estimated for individuals with the relevant qualifications. Thanks to these estimated scores, the test performance of the individual answering the items is determined (Lord & Novick, 2008, p. 359). Item response theory is based on two basic structures:

> The latent traits or competencies of individuals can be identified by the performance of respondents on test items.
>
> The relationship between the competencies of the individuals answering the items and their responses to the items can be expressed by a non-linear function called the item characteristic function (Hambleton et al., 1991, p. 110).

The most important difference between item response theory and classical test theory is that in CTT, ability levels are ignored, and a common estimate of measurement precision is used, which is assumed to be equal for all individuals, whereas in IRT, the latent ability value affects measurement precision (Jabrayilov et al., 2016).

To perform DIF analyses based on IRT, unidimensionality and local independence assumptions must be met, and model data fit must be ensured. Unidimensionality is the measurement of a single latent ability of the items included in the test (Hambleton & Swaminathan, 2013, p. 16). Local independence is explained in the form that the item scores of the study group consisting of individuals with the same ability level are independent of each other (Lord & Novick, 2008, p. 361). There are many IRT models available. The widely used unidimensional IRT models are distinguished from each other according to the number of item parameters, and these models are named according to the number of those parameters (Hambleton et al., 1991, p. 12). Logistics models are divided into three: the one-parameter logistic model (1PL), the two-parameter logistic model (2PL), and the three-parameter logistic model (3PL). In the one-parameter logistic model (1PL), only the item difficulty parameter is estimated (Hambleton et al., 1991, p. 13). In the two-parameter logistic model (2PL), the item discrimination parameter is estimated in addition to the item difficulty parameter (Hambleton et al., 1991, p. 15). In the three-parameter logistic model (3PL), the chance parameter is estimated in addition to the item discrimination and item difficulty parameters (Hambleton et al., 1991, p. 17).

There are many methods to perform DIF analyses based on IRT. Three of the methods mentioned below were used in this study.

Lord's χ^2 method: In Lord's χ^2 method, the variance and covariance of the focus and reference groups are calculated to detect DIF, and these values are scaled for comparison. These scaled values are calculated using Lord's χ^2 method. At the next stage, the null hypothesis is tested by comparing it with critical values and it is decided whether the difference exists (Cromwell, 2002). According to Lord's χ^2 method, the fact that there is a difference between the focus and reference group item parameters of an item indicates that the item functions in a different way. In other words, for an item to contain DIF, the probabilities of individuals with the same ability level in different groups to respond correctly to the relevant item must differ (Kim et al., 1994).

Raju's area measurement method: In Raju's area measurement method, it is checked whether the area value between the item characteristic curves of two different groups at the same ability level is different from zero, or whether the curves overlap. If the curves overlap, in other words, if the area value measured between the curves is zero, it indicates that the item does not contain DIF (Başusta, 2013). The fact that there is an area between the ICC indicates that the item works differently for the two groups and that the item contains DIF (Raju, 1990).

Item Response Theory Likelihood Ratio method: In this method, the hypothesis of a difference between focus and reference group item parameters is checked. In this respect, restricted and generalized models are created, and the ratios of these models are compared (Atalay et al., 2012). In other words, the significance of the differences in the likelihood ratio values obtained to determine the model-data fit of the restricted and generalized model is tested (Thissen, 2001). The restricted model assumes that the item parameters are the same for the reference and focus groups. In contrast, in the generalized model, it is assumed that the parameters of an item are different for each group while the parameters for the other items are equal. The restricted model is created separately for each item in the study and proportioned to the extended model (Gök et al., 2014).

As with every statistical analysis finding, DIF findings are also affected by the structure and characteristics of the data, such as missing data and outliers.

Missing data is defined as the difference between the data intended to be observed and the observed data (Longford, 2005, p. 13). There are many reasons for the occurrence of missing data. For example, missing data may exist due to some individuals in the sample not answering the questions unconsciously, participants preferring not to answer the questions, participants leaving the study before it is completed, problems arising during data collection or problems arising from the data collection tool, and due to errors made during data entry (Osborne, 2013, pp. 106-108).

Missing data can cause some problems: It can create bias in the estimations in statistical analyses, reduce the power of the analysis and lead to higher standard error values due to lack of information. Furthermore, frequently used statistical methods cannot be applied to datasets with missing data leading to improper use of assessable resources (Peng et al., 2006).

In order to make accurate predictions in research, a solution to the problem of missing data should be found before proceeding with the analysis. In this direction, researches may consider solutions such as including new values in the data, not including cells with missing data in the dataset, making predictions about missing data and imputing approximate values instead of missing data (Çüm & Gelbal, 2015).

To impute values to missing data, it is necessary to choose the appropriate imputation method. For this purpose, firstly, the structure of the missing data is examined, and the appropriate imputation method is selected. Missing data can occur in three different mechanisms: missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR):

MCAR is defined as missing data that is not affected by the variable in which the missing data is located and is not caused by other variables such as language or gender (Çüm et al., 2018). For example, if the missing values in a dataset consisting of students' answers to exam questions do not differ for students with high or low scores, or if any other variable did not have an effect on the missing values, it can be said that the missing data are distributed completely at random (MCAR).

MAR means that the missing data for a variable are not caused by that variable but by the effect of one or more other variables in the model (Enders, 2010, p. 6). For example, the fact that the missing data in the variable consisting of students' answers to the exam questions do not differ according to the high or low scores obtained, but the effect of one or more other variables on the losses shows that the missing data are randomly distributed.

MNAR is defined as the probability that having missing data in a variable is related to the values of the relevant variable even after controlling other variables. In other words, the probability of missing data affects the variable with missing values (Enders, 2010, p. 8). For example, the fact that the missing values obtained from the variable including students' answers to the exam questions differ for individuals with high or low scores but are not affected by other variables shows that missing data are distributed not at random (MNAR).

There are methods suitable for missing data mechanisms. Among these methods, the ones used in the research are explained below:

Regression Imputation (RI) Method: In the regression imputation method, a regression equation is first established that predicts the missing data from the complete data. Then, estimated values are created, and these values are substituted for the missing data to obtain a complete dataset (Enders, 2010, p. 44). Regression imputation provides unbiased parameter estimates in MCAR and MAR missing data mechanisms (Baraldi & Enders, 2010). This method has some negative features: Since the missing data are estimated based on the complete data, results close to the other data will be obtained. Therefore, results similar to the real data will not be obtained. And the variance will decrease because the data obtained by regression imputation make predictions close to the average. When the independent variables are not good, this method will reach the same results as the mean imputation method because it will not be able to predict the missing data accurately. Finally, this method cannot be used when the value obtained with the regression imputation method is not within the data value range (Tabachnick & Fidell, 2013, p. 68).

Multiple imputation (MI) Method: In this method, the missing data imputation process takes place in three steps. In the first step, m (m>1) complete datasets are created. In the second step, m different datasets are analysed with standard methods. Finally, the results of the analyses are combined to form a single dataset (Schafer & Graham, 2002). In this method, missing data imputation is iterated at least 2 times and there is no limit to the number of iterations. A large number of imputations with the MI method reduces the standard error (Schafer & Olsen, 1998). This method makes accurate inferences even in MAR and MNAR mechanisms (Van Buuren, 2018, p. 48).

K-Nearest Neighbor Method (kNN): In this method, data imputation is performed by distance-based classification (Cihan, 2018). The kNN method imputes missing data in four stages. In the first stage, the distances between the target data and other data are calculated. In the second stage, these distances are ranked, and in the third stage, the k smallest values between the ranked distances are taken. In the last stage, the target data is imputed to the most repeated class among the k values (Altay, 2016). The characteristics of all groups should be identified in advance. The effectiveness of the k-nearest neighbor method is affected by some conditions. The number of neighbors, threshold value, similarity measurement and sufficient number of normal actions in the learning set can be given as examples (Çalışkan & Soğukpınar, 2008).

Like many statistical methods, DIF analyses are also affected by the existence of missing data since they are developed for complete datasets. Therefore, if there is missing data in the dataset, the missing data problem should be solved with appropriate methods and the dataset should be made complete before proceeding to DIF analysis. A review of the literature reveals that there are few studies on the effect of missing data imputation methods on DIF. In the studies examining the effect of missing data imputation methods on DIF, it has been found out that DIF methods based on CTT are generally used or DIF methods based on CTT are compared with DIF methods based on IRT (Dinçsoy, 2022; Emenogu et al., 2010; Garrett, 2009; Robitzsch & Rupp, 2009; Selvi & Alıcı, 2018; Tamcı, 2018). The fact that DIF methods based on IRT are not generally used in the studies revealed that conducting a study on DIF methods based on IRT would contribute to the literature. At the same time, because of this study, it is aimed to enable the selection of appropriate imputation methods for future IRT-based DIF analyses. Based on this objective, this study examines the effects of regression imputation (RI), multiple imputation (MI), k-nearest neighbor (kNN) methods on DIF detection through Lord's χ2, Raju's area measurement, item response theory likelihood ratio methods.

2. METHOD

2.1. Research Model

This research aims to examine the effect of regression imputation (RI), multiple imputation (MI), and *k*-nearest neighbor (kNN) methods on DDIF to deal with missing data in a dataset containing missing values at different rates considering the variables of language and gender using Lord's χ^2, Raju's area measurement and item response theory likelihood ratio methods. For this reason, a descriptive survey model was used in the study. The descriptive survey model examines existing phenomena in terms of conditions, practices, beliefs, processes, relationships, or trends (Salaria, 2012).

2.2. Study Group

International studies such as the Programme for International Student Assessment (PISA) conducted by the Organisation for Economic Cooperation and Development (OECD) allow comparing the performance of students in different countries (Taş et al., 2016). To carry out DIF analyses in terms of the language variable in the study, PISA 2018 data from different countries were used in this study. Accordingly, the study group of the research consists of students who answered the PISA 2018 science literacy test. 600.000 students participated in PISA 2018, representing approximately 32 million students in the 15-age group from 79 participating countries and economies (OECD, 2019). For the study, Türkiye and the countries that use English as their mother tongue, one of the languages in which the PISA 2018 tests were developed, were selected to conduct DIF analyses for the language variable. In the selection of the countries whose mother tongue is English, attention was paid to pick the ones with the closest science averages to each other. For this reason, the United Kingdom, the United States, New Zealand, and Australia were included in this study. In PISA 2018, 6.890 students from Türkiye, 13.818 students from the United Kingdom, 4.838 students from the United States, 6.173 students from New Zealand, and 14.273 students from Australia participated (OECD, 2019).

Sample sizes with equal focal and reference groups reduce the error rates of DIF detection methods based on IRT (Sünbül & Sünbül, 2016). Thus, as many native English speakers as the number of native Turkish speakers were included in the analysis through simple random sampling. There were 530 students from Türkiye who participated in PISA 2018 answering booklets 14 and 15. Since 300 students out of 530 students had the complete responds, those 300 students from Türkiye were included in the study. Accordingly, out of 1.756 students from native English-speaking countries who answered the items in booklets 14 and 15 and had complete respondst, 300 of them were chosen for the analysis by simple random sampling method. A total of 600 students from Türkiye, the United Kingdom, the United States, New Zealand, and Australia were included in the analysis. When the literature was examined, it was seen that the sample size of the focus and reference groups should be larger than 200 in the analyses related to DIF because it is important in obtaining accurate results (Jodoin & Gierl, 2001; Rogers & Swaminathan, 1993). Based on this, it can be stated that accurate DIF results can be obtained from the analyses when the sample is examined. Table 1 shows the distribution of individuals in the study group by country and gender with science test means of countries.

Table 1. *Distribution of individuals in the study group by country and gender and science means of countries.*

Country	Female	Male	Total	Mean Science Literacy
Australia	41	70	121	503
United Kingdom	37	53	90	505
New Zealand	18	23	41	508
Türkiye	147	153	300	468
USA	22	26	48	502
Total	283	317	600	

When Table 1 is examined, there are 121 students (41 female and 70 male) from Australia in the dataset. There are 90 students (37 female and 53 male) from the United Kingdom; 41 students (18 female and 23 male) from New Zealand; 300 students (147 female and 153 male) from Türkiye, and 48 students (22 female and 26 male) from the USA. The dataset of the study consisted of 600 students comprising 283 females and 317 males. When the mean science literacy scores of the countries are analysed, Australia has a mean score of 503, the United Kingdom 505, New Zealand 508, Türkiye 468, and the USA 502.

2.3. Data Collection Tools

This study was conducted on booklets numbered 14 and 15, which have the highest number of common items among the booklets used for Turkish and English languages in the PISA 2018 application and which provide content validity. The booklets included in the study had a total of 20 common items, 5 open-ended and 15 multiple-choice items. Correct answers were coded as "1" and incorrect answers were coded as "0". In the items with partially correct answers, incorrect answers were coded as "1", partially correct answers as "11" and "12", and correct answers as "21". The answers with the codes "5, 6, 7, 8, 9, 96, 97, 98, 99" were included in the analysis with the missing data code as in the PISA 2018 codebook. In the item numbered DS657Q04C with partially correct answers, answers coded "21" were coded as "1"; answers coded "1", "11" and "12" were coded as "0" and included in the analysis. The data for PISA 2018 were published on the OECD website in 2019 (https://www.oecd.org/pisa/data/). Table 2 provides information about the items included in the study.

Table 2. *Science literacy items used in the analysis.*

Item	Unit	Scientific competencies	Content
CS466Q01S	Forest Fires	Evaluate and design scientific enquiry	Physical
CS466Q07S	Forest Fires	Evaluate and design scientific enquiry	Physical
CS256Q01S	Spoons	Explain phenomena scientifically	Physical
DS326Q01C	Milk	Interpret data and evidence scientifically	Living
DS326Q02C	Milk	Interpret data and evidence scientifically	Living
CS326Q03S	Milk	Interpret data and evidence scientifically	Living
CS326Q04S	Milk	Interpret data and evidence scientifically	Physical
CS602Q01S	Urban Heat Island Effect	Interpret data and evidence scientifically	Earth and Space
CS602Q02S	Urban Heat Island Effect	Explain phenomena scientifically	Earth and Space
DS602Q03C	Urban Heat Island Effect	Explain phenomena scientifically	Physical
CS602Q04S	Urban Heat Island Effect	Interpret data and evidence scientifically	Living
CS603Q03S	Elephants and Acacia Trees	Explain phenomena scientifically	Living
DS603Q02C	Elephants and Acacia Trees	Evaluate and design scientific enquiry	Living
CS603Q03S	Elephants and Acacia Trees	Explain phenomena scientifically	Living
CS603Q03S	Elephants and Acacia Trees	Explain phenomena scientifically	Living
CS603Q05S	Elephants and Acacia Trees	Evaluate and design scientific enquiry	Living
CS657Q01S	Invasive Species	Explain phenomena scientifically	Living
CS657Q02S	Invasive Species	Explain phenomena scientifically	Living
CS657Q03S	Invasive Species	Interpret data and evidence scientifically	Living
DS657Q04C	Invasive Species	Explain phenomena scientifically	Living

When Table 2 is examined, it can be observed that the PISA 2018 science literacy test items included in the study are found in the units of forest fires, spoons, milk, urban heat island effect, elephants and acacia trees, and invasive species. The items measure the skills of evaluating and designing scientific research, explaining phenomena scientifically, and interpreting data and evidence scientifically. Physical, living, Earth and Space titles constitute the content areas of the items.

2.4. Data Analysis

In the study, outliers and descriptive statistics were checked first via the IBM SPSS 26.0 program. Then, confirmatory factor analysis was conducted with the "lavaan" package of the R Studio program to test the unidimensionality and local independence assumptions regarding IRT (Rosseel et al., 2017). R Studio program "ltm" package was used to examine model-data fit (Rizopoulos & Rizopoulos, 2018). The population heterogeneity of the dataset was examined with the "Equaltest MI" package of the R Studio program (Jiang et al., 2022). After reviewing the suitability of the dataset for analysis, four datasets with 5%, 10%, 20% and 30% of missing data suitable for the MCAR mechanism were created from the complete dataset with the R Studio program "MissMethods" package (Josse et al., 2022) and the missing data mechanisms of the datasets were checked with the IBM SPSS 26.0 program. In the following stage, the missing data were imputed via the RI and MI methods using the IBM SPSS 26.0 program and the kNN method using the R Studio program "VIM" (Templ et al., 2016) package. With the MI method, imputations were made by selecting 5 as the imputation number and 5 different datasets belonging to each missing data rate were obtained. For each dataset with missing data rates in the study, the average of the DIF analyses of the 5 imputations made with the MI method were combined in a common DIF result. DIF analyses were performed with Lord's χ^2, Raju's area measurement and item response theory likelihood ratio methods using the R Studio program "difR" (Magis et al., 2015) package in terms of gender and language variables for the datasets completed by imputations via RI, MI, kNN methods. The values obtained from the complete datasets were taken as a reference and compared with the results obtained from the datasets in which missing data were imputed.

2.4.1. *Outliers*

Outliers are explained as data with values outside the usual values or extreme values (Çokluk et al., 2021, p. 2). Outliers can occur in two ways: univariate and multivariate. Univariate outliers can be detected by statistical methods such as converting all raw scores in the distribution into standard Z scores. For a subject to be an outlier, the Z value must be less than -3 and greater than +3 (Çokluk et al., 2021, p.14). To detect the univariate outliers in the dataset, the Z values were examined. As a result of the analysis conducted to detect the Z value, it was found that there are no univariate outliers in the dataset since a Z value less than -3 and greater than +3 was not detected. To determine the multivariate outliers, Mahalanobis Distance, which measures a single data distance from the center or sample mean in the space of the independent variable, is used. A Mahalanobis Distance value of p<0.001 indicates that multivariate outliers are present in the dataset (Çokluk et al., 2021, p.15). When the Mahalanobis Distance was examined for the dataset, the data with a value less than 0.001 could not be determined and it was seen that the multivariate outliers were not present in the dataset.

2.4.2. *Descriptive test statistics*

Some statistical options such as kurtosis and skewness coefficients can be used to assess the normality of the dataset. Skewness and kurtosis coefficients between +1 and -1 indicate that the group does not deviate excessively from the normal distribution (Çokluk et al., 2021, p. 16).

In this study, internal consistency was tested by examining the Kuder Richardson-20 (KR-20) coefficient. A KR-20 reliability coefficient of 0.70 and above indicates that the internal consistency value is at an acceptable level (De Vellis, 2003, p. 95).

In this section, the normality of the data was examined. Table 3 presents the findings related to the normality and reliability tests.

Table 3. *Test statistics, normality tests and reliability coefficients related to sub-problems.*

Statistics	Gender		Language	
	Female	Male	Turkish	English
Number of Students	283	317	300	300
Mean	11.6	11.82	11.02	12
Median	12	12	11	13
Mode	9	15	11	15
Standard Deviation	3.96	4.11	3.97	3.99
Range	18	20	18	20
Skewness	-.0.31	-0.36	-0.08	-0.63
Kurtosis	-0.60	-0.55	-0.63	-0.19
KR-20	0.77	0.79	0.76	0.79

When examining Table 3, it is evident that the measures of central tendency are relatively close to each other. Skewness and kurtosis coefficients are in the range of +1 and -1. This indicates that the distribution is close to normal (Çokluk et al., 2021, p. 16). The KR-20 reliability coefficients of 0.70 and above in all groups indicate that the reliability principle of the groups is met.

2.4.3. *Confirmatory factor analysis*

In this study, confirmatory factor analysis was performed on the complete dataset with the R Studio program “lavaan” package to examine whether the data has met the unidimensionality assumption (Rosseel et al., 2017).

Table 4. *Confirmatory factor analysis model data fit values.*

Indices	Value
SBχ^2	222.31
Degrees of freedom	167
RMSEA	0.02
SRMR	0.03
TLI	0.94
CFI	0.95

As a result of confirmatory factor analysis, SBχ^2, degrees of freedom, RMSEA, SRMR, TLI, and the CFI values were obtained and the unidimensionality assumption was checked based on these values. The Tucker and Lewis index (TLI) value above 0.97 indicates perfect fit, above 0.95 indicates very good fit, and above 0.85 indicates acceptable fit. The standardized root mean square of residuals (SRMR) values close to 0 are considered excellent and values less than 0.05 are considered good. The root mean square error of approximation (RMSEA) value is considered good when it is 0.05 and less, acceptable between 0.05 and 0.08, and poor when it is 0.10 and above. The comparative fit index (CFI) shows an acceptable fit between 0.95 and 0.97 (Erdoğan, 2019). Based on this information, when Table 4 created as a result of confirmatory factor analysis is examined, it is determined that all values provide model-data fit. This shows that the unidimensionality assumption is met.

Local independence is an assumption related to the unidimensionality assumption. If the unidimensionality assumption is met in a test, the items in the test also meet the local independence assumption (Hambleton & Swaminathan, 2013, p. 23). Accordingly, it can be stated that the items in the study meet the local independence assumption.

2.4.4. *Population heterogeneity*

In this study, to determine the suitability of the dataset for the analysis, the population heterogeneity of the dataset was checked in terms of language and gender variables using the "Equaltest MI" package of the R Studio program (Jiang et al., 2022). To determine population heterogeneity, Model 5 and Model 6 were compared for equality in latent means. S-Bχ^2(*df*), χ^2/*df*, $\Delta\chi^2$(Δdf), RMSEA, ΔRMSEA goodness-of-fit indices of the models were taken into account during the comparison. A value range of $0 \leq \chi^2$ /*df*≤ 2 indicates a good fit and a value range of $2 \leq \chi^2$ /*df*≤ 3 indicates an acceptable fit. While a value range of 0≤RMSEA≤0.05 indicates a good fit, and a value range of 0.05≤RMSEA≤0.08 indicates an acceptable fit (Schermelleh-Engel et al., 2003). In this study, the change between Model 5 and Model 6 was evaluated by considering ΔCFI≤0.01 and ΔRMSEA ≤0.01 (Taşkıran & Şenel, 2022).

Table 5. *Population heterogeneity fit indices of the dataset by language and gender variables.*

	Model	S-B χ^2(df)	χ^2 /df	$\Delta\chi^2$ (Δdf)	CFI	ΔCFI	RMSEA	ΔRMSEA
Language	Model 5	618.41 (388)	1.75		0.77		0.05	
	Model 6	702.62 (391)	1.80	21.21(3)	0.75	0.01	0.05	0.00
Gender	Model 5	455.78 (388)	1.17		0.95		0.02	
	Model 6	458.91 (391)	1.17	3.12(3)	0.95	0.00	0.02	0.00

p<0.05, Model 5 = Equality of variance, Model 6 = Equality of Latent Means

When the χ^2 /*df* indexes are examined in terms of the language variable in Table 5, the fact that Model 5 has a value of 1.75 and Model 6 has a value of 1.80 indicates that both models show a good fit. The ΔRMSEA value of 0 indicates that this fit index is at an acceptable level. Based on this, it can be said that there is a good fit between the models. When the ΔCFI fit index is examined, the fact that this value is 0.01 indicates that the fit index is at an acceptable level proving that there is a good fit between the models.

Considering the χ^2 /*df* index in terms of the gender variable, Model 5 and Model 6 have a value of 1.17 indicating a good fit. ΔRMSEA value of 0 indicates that the fit index is at an acceptable level and there is a good fit between the models. A ΔCFI value of 0 indicates that the fit index is at an acceptable level and there is a good fit between the models. According to the results of the population heterogeneity analysis, it was determined that there was no difference between the latent means for both variables.

2.4.5. *Model-data fit*

In this study, model-data fit was examined through the "ltm" package in the R Studio program (Rizo-Poulos & Rizopoulos, 2018). For this reason, the likelihood ratio test (logLik), Akaike information criterion (AIC) and Bayesian information criterion (BIC) values were compared, and the p-value and degrees of freedom obtained as a result of ANOVA were analyzed. Table 6 shows the results of the model-data fit analysis.

Table 6. *Model data fit comparison.*

Model	logLig	AIC	BIC	Number of Parameters	degrees of freedom	*p*
Rasch-1PL	-6681.39	13404.78	13497.11	14		
2PL	-6632.76	13345.53	13521.40	14	19	0
3PL	-6620.33	13360.67	13624.48	16	20	0.20

When Table 6 is examined, the fact that the *p*-value of the 3PL model is not significant (*p*>0.05) indicates that the model is not suitable for analysis. The fact that the loglik and AIC values of the 2PL model are smaller than the loglik and AIC values of the 1PL model indicates that the 2PL model is suitable for the study. Although the fact that the BIC value of the 1PL model is less than the 2PL model does not support this situation, the fact that the variance analysis result of the 2PL model is significant shows that 2PL model fits better than other models and as a result, the 2PL model is the appropriate model for the analysis. In the study, after factor analysis, population heterogeneity and model-data fits were examined, four datasets with 5%, 10%, 20% and 30% of missing data suitable for the MCAR mechanism were created from the complete dataset and the missing data mechanisms of the datasets were checked. In the next stage, the datasets were completed by imputing missing data using RI, MI and kNN methods. DIF analyses were performed on the newly obtained datasets with gender and language variables using Lord's χ^2 method, Raju's area measurement method and item response theory likelihood ratio method. As a result of the analyses, items with a *p*-value below 0.05 and DIF finding in two of the three DIF methods were accepted to contain DIF. Accordingly, DIF analyses were performed on the complete dataset and datasets with missing data imputation. The values obtained from the complete datasets were compared with the results obtained by data imputation.

3. FINDINGS

In this section, the results of the DIF analyses are presented. In the analyses, Lord's χ^2 method, Raju's area measurement method, and item response theory likelihood ratio method are applied for gender and language variables. The analyses were carried out on the complete dataset and the one with missing data. The missing dataset was completed by imputing 5%, 10%, 20%, and 30% via RI, MI, and kNN methods. In Table 7, the DIF results obtained by Lord's χ^2 method, Raju's area measurement method and item response theory likelihood ratio method from the complete dataset and the datasets completed by imputing 5%, 10%, 20% and 30% in terms of the language variable and Table 8 in terms of the gender variable are compared. If at least two of the three DIF detection methods used in the study showed DIF, the related item was considered to contain DIF. In Table 7 and Table 8, in the complete dataset and in the datasets completed with RI, MI, and kNN methods at the rates of 5%, 10%, 20%, and 30%, "DIF" was written in front of the items that showed DIF in at least two DIF detection methods and it was stated that the relevant item contained DIF.

In Table 7, the items in the complete dataset and the datasets completed with RI, MI and kNN methods at 5%, 10%, 20% and 30% of rates were identified as DIF items in terms of the language variable using Lord's χ^2 method, Raju's area measurement method, and item response theory likelihood ratio method. If DIF was identified in at least two methods among the items in the datasets, it was accepted that the item contained DIF. Accordingly, DIF was detected in 6 items (CS256Q01S, CS326Q04S, CS602Q01S, CS603Q01S, DS603Q02C, CS603Q03S) out of 20 items included in the analysis in the complete dataset.

DIF was detected in 6 items (CS256Q01S, CS326Q04S, CS602Q01S, CS602Q02S, CS603Q03S, CS603Q04S) in the dataset that was imputed at 5% with the RI method. There was a 67% agreement between the complete dataset and the dataset completed by 5% with the RI method regarding items containing DIF.

In the dataset completed 10% by the RI method, DIF was detected in 5 items (CS256Q01S, CS326Q04S, CS602Q01S, S603Q02C, CS603Q03S). Based on this, 83% agreement was found between the items with DIF in the complete dataset and those with DIF in the dataset completed 10% with the RI method.

In the dataset with 20% missing data imputation by the RI method, DIF was found in 3 items (DS603Q02C, CS603Q03S, CS603Q04S). Between the complete dataset and the dataset completed by the RI method at the rate of 20%, the rate of the same items containing DIF was determined as 33%.

In the dataset completed 30% with the RI method, DIF was found in 2 items (CS603Q03S, CS603Q05S). The probability of the same items containing DIF was found to be 17% in the dataset in which 30% of the data were imputed by the RI method.

DIF was detected in 6 items (CS326Q04S, CS602Q01S, CS603Q01S, DS603Q02C, CS603Q03S, CS603Q04S) in the dataset completed 5% with the MI method. It was observed that 83% of the items with DIF in the complete dataset also contained DIF in the one completed 5% with the MI method.

Table 7. *Findings of item response theory-based differential item functioning (Lord's χ^2, Raju's area measurement, item response theory likelihood ratio) analysis of complete dataset and datasets with different ratios of missing data and completed with different imputation methods (regression imputation, multiple imputation and k-nearest neighbor method) in terms of the language variable.*

	DIF Status												
	RI Method					MI Method				kNN Method			
Item	complete dataset	5%	10%	20%	30%	5%	10%	20%	30%	5%	10%	20%	30%
CS466Q01S													
CS466Q07S													
CS256Q01S	DIF	DIF	DIF							DIF			
DS326Q01C								DIF					
DS326Q02C									DIF				
CS326Q03S									DIF			DIF	
CS326Q04S	DIF	DIF	DIF			DIF	DIF	DIF	DIF	DIF	DIF	DIF	DIF
CS602Q01S	DIF	DIF	DIF			DIF	DIF	DIF	DIF	DIF	DIF	DIF	DIF
CS602Q02S		DIF											
DS602Q03C													
CS602Q04S													
CS603Q01S	DIF					DIF							
DS603Q02C	DIF		DIF	DIF		DIF		DIF		DIF		DIF	
CS603Q03S	DIF	DIF	DIF	DIF	DIF	DIF	DIF	DIF	DIF	DIF	DIF	DIF	DIF
CS603Q04S		DIF		DIF		DIF						DIF	DIF
CS603Q05S					DIF								
CS657Q01S												DIF	
CS657Q02S													DIF
CS657Q03S													
DS657Q04C									DIF				

DIF was detected in 3 items (CS326Q04S, CS602Q01S, CS603Q03S) in the dataset that was made complete by imputing 10% of data with the MI method. 50% of the items with DIF in the complete dataset also showed DIF in the dataset with 10% of data imputation by the MI method.

In the dataset, where 20% of the missing data was imputed with the MI method, DIF was observed in 5 items (DS326Q01C, CS326Q04S, CS602Q01S, DS603Q02C, CS603Q03S). 67% of the items detected DIF in the complete dataset contain DIF in the dataset with 20% of data imputation by the MI method.

In the dataset completed 30% with the MI method, DIF was detected in 6 items, including items numbered DS326Q02C, C6S326Q03S, CS326Q04S, CS602Q01S, CS603Q03S, DS657Q04C. 50% of the items containing DIF in the complete dataset also contain DIF in the one completed 30% with the MI method.

It was observed that 5 items (CS256Q01S, CS326Q04S, CS602Q01S, DS603Q02C, CS603Q03S) contained DIF in the dataset with 5% of imputation by the kNN method. It was found that the items containing DIF in the dataset completed by the kNN method at the rate of 5% were the same items as 83% of the items detected DIF in the complete dataset.

In the dataset completed 10% applying the kNN method, DIF was detected in 3 items: CS326Q04S, CS602Q01S, and CS603Q03S. 50% of the items with DIF in the full data set also showed DIF in the data set where 10% were assigned by the MP method.

In the dataset with 20% missing data imputation by the kNN method, DIF was detected in 7 items (C6S326Q03S, CS326Q04S, CS602Q01S, DS603Q02C, CS603Q03S, CS603Q04S, CS657Q01S). The ratio of the number of common items between the items containing DIF in the complete dataset and the items containing DIF in the dataset in which 20% of the data was imputed with the kNN method is 67%.

5 items (CS326Q04S, CS602Q01S, CS603Q03S, CS603Q04S, CS657Q02S) contain DIF in the dataset completed by imputing 30% with the kNN method. 50% of the items with DIF in the complete dataset are compatible with the dataset made 30% complete by the kNN method.

Table 8. *Findings of item response theory-based differential item functioning (Lord's χ^2, Raju's area measurement, item response theory likelihood ratio) analysis of complete dataset and datasets with different ratios of missing data and completed with different imputation methods (regression imputation, multiple imputation and k-nearest neighbor method) in terms of the gender variable.*

	DIF Status												
	RI Method					MI Method				KNN Method			
Item	complete dataset	5%	10%	20%	30%	5%	10%	20%	30%	5%	10%	20%	30%
CS466Q01S													
CS466Q07S													
CS256Q01S													
DS326Q01C													
DS326Q02C				DIF								DIF	
CS326Q03S												DIF	
CS326Q04S													
CS602Q01S													
CS602Q02S													
DS602Q03C													
CS602Q04S													
CS603Q01S	DIF		DIF							DIF	DIF		
DS603Q02C													
CS603Q03S													
CS603Q04S													
CS603Q05S													
CS657Q01S													
CS657Q02S													
CS657Q03S													

In Table 8, items showing DIF in terms of the gender variable were identified through Lord's χ^2 method, Raju's area measurement method, and item response theory likelihood ratio method from the items in the complete dataset and the datasets completed with RI, MI and kNN methods at 5%, 10%, 20% and 30% of rates. If DIF was detected in at least two methods, it was accepted that the item contained DIF. Based on this, DIF was found in the item DS603Q01S included in the analysis of the complete dataset.

DIF could not be determined in any item in the datasets completed by imputing 5% and 30% of missing data using the RI method. This shows that the DIF inclusion rate of the same items is 0% between the datasets with 5% and 30% of data imputation using the RI method and the complete dataset.

The detection of DIF in the item CS603Q01S in the dataset completed at the rate of 10% by RI shows that the same item contains DIF both in the complete dataset and the dataset imputed 10% by the RI method. DIF inclusion rate of the same items is 100% between the complete dataset and the one with %10 data imputation using the RI method.

In the dataset, completed by imputing the missing data by the RI method at the rate of 20%, DIF was found in item DS326Q02C. This shows that the DIF inclusion rate of the same items is 0% between the dataset with 20% data imputation using the RI method and the complete dataset.

DIF could not be determined in any item completed 5%, 10%, 20% and 30% by the MI method. The fact that there are no items containing DIF in the datasets completed 5%, 10%, 20% and 30% with the MI method indicates that the DIF rate of the complete dataset and these datasets is 0% for the same items.

In the datasets, completed by imputing 5% and 10% by the kNN method, it was found that the item CS603Q01S contained DIF. DIF inclusion rate of the same items is 100% between the datasets with 5% and 10% data imputation using the kNN method and the complete dataset.

DIF was detected in 2 items (DS326Q02C, C6S326Q03S) in the dataset completed by the kNN method at the rate of 20%. The DIF rate of the same items is 0% between the complete dataset and the dataset with %20 data imputation using the kNN method.

It was determined that DIF was not observed in any item in the dataset in which 30% of the missing data were imputed by the kNN method and that different DIF findings were obtained with the complete dataset. There was a 0% agreement between the complete dataset and the dataset completed 30% by the kNN method.

4. DISCUSSION and CONCLUSION

In this study, we examined how DIF results obtained with Lord's χ^2, Raju's area measurement and item response theory likelihood ratio methods change according to the missing data rate using the language and gender variables in the datasets completed by imputing 5%, 10%, 20% and 30% of data using RI, MI and kNN methods. In this regard, the study was conducted on PISA 2018 science literacy test items.

As a result of the analyses, it can be stated that the use of different languages by the individuals responding to the relevant items increases the probability of the items containing DIF because DIF was observed in 6 out of 20 items in the complete dataset regarding the language variable. Observing DIF in 1 out of 20 items in terms of the gender variable in the complete dataset, it can be said that the gender of individuals affects the probability of DIF. With the RI method, the closest result to the complete dataset using the language variable was obtained at a rate of 10%. While better results were obtained at 5% compared to 20% and 30%, the worst result was obtained at 30%. By the gender variable in the completed datasets with the RI method, accurate results were obtained at 10%, while inaccurate results were obtained at 5%, 20% and 30%. In the MI method, the closest result to the complete dataset was obtained at 5% in terms of the language variable while more accurate predictions were made at 20% compared to 10% and 30%. Tamcı (2019) suggested that the MI method should be used when the missing data rate is high. Dinçsoy (2022) found that the MI method was successful in detecting DIF at 10% and 20% of missing data. In the MI method, inaccurate results were obtained at 5%, 10%, 20% and 30% with the gender variable. With the kNN method, values close to the complete dataset were obtained at 5% by the language variable while the most accurate results were obtained after 5% at 20%. DIF was poorly predicted at 10% and 30% rates compared to other rates. The kNN method obtained accurate results at 5% and 10% of missing data rates regarding the gender variable, but inaccurate results were obtained at 20% and 30% of rates. Based on the results of the study, it can be said that the RI method can be used to make imputations at a 10% missing data rate in future studies analyzing DIF based on IRT by the variables of language and gender. It can be suggested that the RI method should not be used at 5%, 20% and 30% of missing data rates. In terms of the language variable, it can be recommended that MI and kNN methods can be used at a rate of 5% in DIF analysis based on IRT, but these methods should not be used at 10%, 20% and 30% of missing data rates. Since inaccurate results will be obtained with the MI method at 5%, 10%, 20% and 30% by the gender variable, it may be recommended to prefer different missing data imputation methods. It can be suggested that the kNN method can be used in the dataset with 5% and 10% of missing data for the gender variable, but this method should not be preferred at 20% and 30% rates. Since the sample size was kept constant in this study, missing data imputation methods with different sample sizes can be examined in future studies. In this study, missing data with the MCAR mechanism were used. In future studies, DIF analyses can be performed with missing data with MAR and MNAR mechanisms.

There are some limitations in this study. It is limited to the use of regression imputation, multiple imputation and k-nearest neighbor imputation methods, and IRT-based Lord's χ^2 method, Raju's area measurement method and item response theory likelihood ratio method for DIF identification. Therefore, in future studies, different DIF detection methods based on IRT or CTT, different missing data imputation methods, and the effect of those imputation methods on DIF can be examined.

Acknowledgments

This study is produced from the first author's master's dissertation.

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

Fatma Ünal: Investigation, Resources, Visualization, Software, Formal Analysis, and Writing-original draft. **Hakan Koğar:** Methodology, Supervision, and Validation.

Orcid

Fatma Ünal https://orcid.org/0000-0001-6306-4210
Hakan Koğar https://orcid.org/0000-0001-5749-9824

REFERENCES

Altay, O. (2016). *Genetik ve genetik olmayan faktörlere bağlı olarak Türk hastalarda varfarin dozajını tahmin eden bir uzman sistem geliştirilmesi* [*Improvement of an expert system that predict warfarin dosage in Turkish patients depending on genetic and non-genetic factors]* [Master's dissertation, Fırat University]. Higher Education Institution National Thesis Center. https://tez.yok.gov.tr/UlusalTezMerkezi/tezDetay.jsp?id=W663t01X1WehurHffLL0Q&no=Urx32Vn-YC2f6ufE0L3ZTw

Atalay, K., Gök, B., Kelecioğlu, H., & Arsan, N. (2012). Değişen madde fonksiyonunun belirlenmesinde kullanılan farklı yöntemlerin karşılaştırılması: Bir simülasyon çalışması [Comparing different differential item functioning Methods: A simulation study]. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi (H. U. Journal of Education)*, 43, 270-281. https://dergipark.org.tr/tr/pub/hunefd/issue/7795/102030

Atar, B., Atalay Kabasakal, K., Ünsal Özberk, E.B., Özberk, E.H., & Kıbrıslıoğlu Uysal, N., (2021). *R ile veri analizi ve psikometri uygulamaları [Data analysis and psychometric applications with R]* (3th ed.). Pegem Akademi.

Baraldi, A.N., & Enders, C.K. (2010). An introduction to modern missing data analyses. *Journal of school psychology*, *48*(1), 5-37. https://doi.org/10.1016/j.jsp.2009.10.001

Başusta, N.B. (2013). *PISA 2006 fen başarı testinin madde yanlılığının kültür ve dil açısından incelenmesi [Examination of item bias and language perspective of PISA 2006 science and culture achievement test]* [Doctoral dissertation, Hacettepe University]. Hacettepe University Open Archive. https://www.openaccess.hacettepe.edu.tr/xmlui/bitstream/handle/11655/1766/42cc60c5-40f1-4b78-8c75-cc6d7932416e.pdf?sequence=1&isAllowed=y

Bortolotti, S.L.V., Tezza, R., de Andrade, D.F., Bornia, A.C., & de Sousa Júnior, A.F. (2013). Relevance and advantages of using the item response theory. *Quality & Quantity*, 47, 2341-2360.

Cihan, P. (2018). *Veri madenciliği yöntemleriyle hayvan hastalıklarında teşhis, prognoz ve risk faktörlerinin belirlenmesi [Determination of dlagnosis, prognosis and risk factors inanimal diseases using by diseases using by data mining methods]* [Doctoral dissertation, Yıldız Technical University]. Yıldız Technical University Open Archive. http://dspace.yildiz.edu.tr/xmlui/bitstream/handle/1/13155/7932.pdf?sequence=1&isAllowed=y

Cromwell, S. (2002). A primer on ways to explore item bias. https://eric.ed.gov/?id=ED463307

Çalışkan, S.K., & Soğukpınar, İ. (2008). Kxknn: K-means ve k en yakin komşu yöntemleri ile ağlarda nüfuz tespiti [Kxknn: Penetration detection in networks with k-means and k nearest neighbor methods]. *EMO Yayınları*, 120-24. https://www.emo.org.tr/ekler/8c1874c96244659_ek.pdf

Çokluk, Ö., Şekercioğlu, G., & Büyüköztürk, Ş. (2021). *Sosyal bilimler için çok değişkenli istatistik: SPSS ve LISREL uygulamaları [Multivariate statistics for social sciences: SPSS and LISREL applications]* (6th ed.). Pegem Akademi.

Çüm, S., & Gelbal, S. (2015). Kayıp veriler yerine yaklaşık değer atamada kullanılan farklı yöntemlerin model veri uyumu üzerindeki etkisi [The effects of different methods used for value imputation instead of missing values on model data fit statistics]. *Mehmet Akif Ersoy University Journal of Education Faculty, 1*(35), 87-111. https://dergipark.org.tr/tr/pub/maeuefd/issue/19408/206357

Çüm, S., Demir, E.K., Gelbal, S., & Kışla, T. (2018). Kayıp veriler yerine yaklaşık değer atamak için kullanılan gelişmiş yöntemlerin farklı koşullar altında karşılaştırılması [A comparison of advanced methods used for missing data imputation under different conditions]. *Mehmet Akif Ersoy University Journal of Education Faculty*, (45), 230-249. https://dergipark.org.tr/tr/pub/maeuefd/issue/35179/332605

De Vellis, R.F. (2003). *Scale development: Theory and applications.* Applied Social Research Methods Series. Sage Publications, Inc. https://www.academia.edu/42875983/Scale_Developm_ent_Theory_and_Applications_Second_Edition

Dogan, E., Guerrero, A., & Tatsuoka, K. (2005). Using DIF to investigate strengths and weaknesses in mathematics achievement profiles of 10 different countries. *In annual meeting of the National Council on Measurement in Education (NCME), Montreal, Canada.* https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=a23cbcde509e6d6b9cd664236acc2d585b634578

Dinçsoy, L.B. (2022). *Karma testlerde kayıp verilerin değişen madde fonksiyonuna etkisinin incelenmesi [Investigation of the effect of missing data on differantial item functioning in mixed type tests]* [Master's dissertation, Hacettepe University]. Hacettepe University. https://openaccess.hacettepe.edu.tr/xmlui/bitstream/handle/11655/25949/10440993.pdf?sequence=1&isAllowed=y

Emenogu, B.C., Falenchuk, O., & Childs, R.A. (2010). The effect of missing data treatment on Mantel-Haenszel DIF detection. *The Alberta Journal of Educational Research, 56*(4), 459- 469. https://doi.org/10.11575/ajer.v56i4.55429

Enders, C.K. (2010). *Applied missing data analysis* (1th ed.). The Guilford Publications, Inc. http://hsta559s12.pbworks.com/w/file/fetch/52112520/enders.applied

Erdoğan, K.H. (2019). *Doğrulayıcı faktör analizi ve farklı veri setlerinde uygulanması [Confirmatory factory analysis and application to different datasets]* [Master's dissertation, Applied Sciences University of Isparta]. Higher Education Institution National Thesis Center. https://acikbilim.yok.gov.tr/bitstream/handle/20.500.12812/378756/yokAcikBilim_10284258.pdf?sequence=-1&isAllowed=y

Garrett, P. (2009). *A Monte Carlo study investigating missing data, differential item functioning, and effect size.* Georgia State University. https://scholarworks.gsu.edu/cgi/viewcontent.cgi?article=1034&context=eps_diss

Gök, B., Kabasakal, K.A., & Kelecioğlu, H. (2014). PISA 2009 öğrenci anketi tutum maddelerinin kültüre göre değişen madde fonksiyonu açısından incelenmesi [Analysis of attitude items in PISA2009 student questionnaire in terms of differential item functioning based on culture]. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi*, *5*(1), 72-87. https://doi.org/10.21031/epod.64124

Gültekin, S., & Demirtaşlı, N.Ç. (2020). Comparing the test information obtained through multiple choice, open-ended and mixed item tests based on item response theory. *Elementary Education Online*, *11*(1), 251-251. https://www.ilkogretim-online.org/fulltext/218-1596943363.pdf?1697476130

Hambleton, R.K., & Swaminathan, H. (2013). *Item response theory: Principles and applications.* Springer Science & Business Media.

Hambleton, R.K., Swaminathan, H., & Rogers, H.J. (1991). *Fundamentals of item response theory (Vol. 2).* Sage.

Jabrayilov, R., Emons, W.H., & 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

Jodoin, M.G., & Gierl, M.J. (2001). Evaluating type I error and power rates using an effect size measure with the logistic regression procedure for DIF detection. *Applied Measurement in Education*, *14*(4), 329-349. https://eric.ed.gov/?id=EJ642273

Josse, J., Mayer, I., Tierney, N., & Vialaneix, N. (2022). CRAN task view: Missing data. https://mirror.truenetwork.ru/CRAN/web/views/MissingData.html

Kalaycıoğlu, D.B., & Kelecioğlu, H. (2011). Öğrenci Seçme Sınavı'nın madde yanlılığı açısından incelenmesi [Analysis of attitude items in PISA2009 student questionnaire in terms of differential item functioning based on culture]. *Eğitim ve Bilim, 36*(161), 3-13. http://egitimvebilim.ted.org.tr/index.php/EB/article/view/143/280

Kim, S.H., Cohen, A.S., & Kim, H.O. (1994). An investigation of Lord's procedure for the detection of differential item functioning. *Applied Psychological Measurement*, *18*(3), 217-228. https://doi.org/10.1177/014662169401800303

Longford, N.T. (2005). *Missing data and small-area estimation: Modern analytical equipment for the survey statistician.* Springer.

Magis, D., Beland, S., Raiche, G., & Magis, M.D. (2015). Package 'difR'. https://cran.r-project.org/web/packages/difR/difR.pdf

MEB (2019). *Uluslararası öğrenci değerlendirme programı PISA 2018 ulusal raporu [International student assessment program PISA 2018 national report*]. Ankara: Directorate of Measurement, Evaluation and Testing Services, Ministry of National Education. https://www.meb.gov.tr/meb_iys_dosyalar/2019_12/03105347_pisa_2018_turkiye_on_raporu.pdf

OECD (2019). PISA 2018 results volume I: What students know and can do. OECD Publishing. https://www.oecd.org/education/pisa-2018-results-volume-i-5f07c754-en.htm

Peng, C.Y., Harwell, M.R., Liou, S.M., & Ehman, L.H. (2006). Advances in missing data methods and implications for educational research. In S. S. Sawilowsky (Ed.), *Real Data Analysis* (pp. 31-78).

Raju, N.S. (1990). Determining the significance of estimated signed and unsigned areas between two item response functions. *Applied Psychological Measurement*, *14*(2), 197-207. https://conservancy.umn.edu/bitstream/handle/11299/113559/v14n2p197.pdf?sequence=1

Rizopoulos, D., & Rizopoulos, M.D. (2018). Package 'ltm'. https://cran.stat.unipd.it/web/packages/ltm/ltm.pdf

Robitzsch, A., & Rupp, A.A. (2009). Impact of missing data on the detection of differential item functioning: The case of Mantel-Haenszel and logistic regression analysis. *Educational and Psychological Measurement, 69(1),* 18-34. https://doi.org/10.1177/0013164408318756

Rogers, H.J., & Swaminathan, H. (1993). A comparison of logistic regression and Mantel Henszel procedures for detecting differential item functioning. *Applied Psychological Measurement*, *17*(2), 105-116. https://doi.org/10.1177/014662169301700201

Rosseel, Y., Jorgensen, T.D., Rockwood, N., Oberski, D., Byrnes, J., Vanbrabant, L., Savalei, V., Merkle, E., Hallquist, M., Rhemtulla, M., Katsikatsou, M., Barendse, M., Scharf, F., & Du, H. (*June 17*, 2017). Package 'lavaan'. Version 0.6-18. https://cran.r-project.org/web/packages/lavaan/lavaan.pdf

Salaria, N. (2012). Meaning of the term descriptive survey research method. *International Journal of Transformations in Business Management*, *1*(6), 1-7.

Schafer, J.L., & Graham, J.W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, *7*(2), 147–177. https://doi.org/10.1037/1082-989X.7.2.147

Schafer, J.L., & Olsen, M.K. (1998). Multiple imputation for multivariate missing- data problems: A data analyst's perspective. *Multivariate Behavioral Research*, *33*(4), 545-571. https://doi.org/10.1207/s15327906mbr3304_5

Schermelleh-Engel, K., Moosbrugger, H., & Müller, H. (2003). Evaluating the fit of structural equation models: Tests of significance and descriptive goodness-of-fit measures. *Methods of Psychological Research Online*, *8*(2), 23-74.

Selvi, H., & Alıcı, D. (2018). Investigating the impact of missing data handling methods on the detection of differential item functioning. *International Journal of Assessment Tools in Education*, *5*(1), 1-14. https://files.eric.ed.gov/fulltext/EJ1250131.pdf

Sırgancı, G., & Çakan, M. (2020). Sıralı lojistik regresyon ve poly-sıbtest yöntemleri ile değişen madde fonksiyonunun belirlenmesi [Determination of the differential item function with ordered logistic regression and poly-sibtest methods]. *Abant İzzet Baysal Üniversitesi Eğitim Fakültesi Dergisi, 20*(1), 705-717. https://doi.org/10.17240/aibuefd.2020.20.52925-665084

Sünbül, S.Ö., & Sünbül, Ö. (2016). Değişen madde fonksiyonunun belirlenmesinde kullanılan yöntemlerde I. Tip hata ve güç çalışması [Type I error rates and power study of several differential item functioning determination methods]. *İlköğretim Online*, *15*(3), 882-897. https://doi.org/10.17051/io.2016.10640

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

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]* [Master's dissertation, Hacettepe University]. Hacettepe University Open Archive. https://openaccess.hacettepe.edu.tr/xmlui/handle/11655/5315

Taş, U.E., Arıcı, Ö., Ozarkan, H.B., & Özgürlük, B. (2016). PISA 2015 ulusal raporu [PISA 2015 national report]. *Ministry of National Education.* https://odsgm.meb.gov.tr/test/analizler/docs/PISA/PISA2015_Ulusal_Rapor.pdf

Taşkıran, C., & Şenel, E. (2022). Çok boyutlu sportmenlik yönelimi ölçeğinin ölçme eşdeğerliğinin test edilmesi [Testing the measurement invariance of the multidimensional sportspersonship orientation scale]. *International Journal of Sport Exercise and Training Sciences-IJSETS*, *8*(4), 190-196. https://doi.org/10.18826/useeabd.1156699

Templ, M., Alfons, A., Kowarik, A., Prantner, B., & Templ, M.M. (2016). VIM: Visualization and Imputation ofMissing Values. R package version 4.6.0, URL https://CRAN.R-project.org/package=VIM

Thissen, D. (2001). *IRTLRDIF v.2.0b: Software for the computation of the statistics involved in item response theory likelihood-ratio tests for differential item functioning*. Chapel Hill: L.L. Thurstone

Psychometric Laboratory, University of North Carolina at Chapel Hill.

Uyar, Ş. (2015). *Gözlenen gruplara ve örtük sınıflara göre tanımlananları madde etkilerinin karşılaştırılması [Comparing differential item functioning based on manifest groups and latent classes]* [Doctoral dissertation, Hacettepe University]. Hacettepe University Open Access System. https://openaccess.hacettepe.edu.tr/xmlui/handle/11655/1816

Van de Vijver, F.J., & Tanzer, N.K. (1997). Bias and equivalence in cross-cultural assessment: An overview. *European Review of Applied Psychology*, *47*(4), 263-279. https://pure.uvt.nl/ws/files/225989/26727_11858.pdf

Van Buuren, S. (2018). *Flexible imputation of missing data*. *Chapman & Hall/CRC Press.*

Yılmaz, M. (2021). *Eğilim puanları kullanılarak ABİDE çalışmasındaki maddelerin değişen madde fonksiyonu açısından incelenmesi [Investigation of differantial item functioning of the test items in the abide study by using propensity scores]* [Master's dissertation, Hacettepe University]. Hacettepe University Open Access System. https://openaccess.hacettepe.edu.tr/xmlui/handle/11655/23603