ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
2020; 11(3); 312-324

Investigation of the Effect of Missing Data Handling Methods on Measurement Invariance of Multi-Dimensional Structures *

Mehmet Ali IŞIKOĞLU ** Burcu ATAR ***

Abstract

The purpose of this study was to compare the missing data handling methods on measurement invariance of multi-dimensional structures. For this purpose, data of 10857 students who participated in PISA 2015 administration from Turkey and Singapore and fully responded to the items related to affective characteristics of science literacy was used. Data with different percentages of missing data (5%, 10%, and 20% missing data) were generated from the complete data set with missing completely at random (MCAR) mechanism. In all data sets, missing data was completed with listwise deletion (LD), serial mean imputation (SMI), regression imputation (RI), expectation maximization (EM), and multiple imputation (MI) methods. Measurement invariance of the construct being measured between countries on completed data sets was investigated with multiple-group confirmatory factor analysis (MG-CFA). Findings from each dataset were compared with reference values. In the results of the study, RI and MI methods in the data set with 5% missing, EM method in the data set with 10% missing, and MI method in the data set with 20% missing gave the more similar results to the reference values than the other methods.

Key Words: Missing data handling methods, measurement invariance, multiple-group confirmatory factor analysis, PISA 2015, science literacy.

INTRODUCTION

Measurement instruments are of great importance in education systems. In order to train qualified workforce in accordance with the needs of the society, placement of individuals in educational institutions and programs, making changes and improvements in educational systems can be made based on the findings obtained from measurement instruments. As a result of national and international assessment studies, countries can even change their educational policies. In particular, the results of large-scale assessment studies that enable international comparisons are followed with interest by all stakeholders of the education.

PISA (Program for International Student Assessment) and TIMSS (Trends in International Mathematics and Science Study) aim to make cross-country comparisons. PISA and TIMSS are large-scale studies that aim to make comparisons between countries and can affect educational policies at national and international levels. The comparability of the results, especially in international assessments, is of great importance in the evaluation of countries. To be able to interpret the findings from different groups who took the same measurement instrument, the measurement instrument should have the same meaning for all groups. In this context, the concept of measurement invariance emerges. Drasgow (1984) defines measurement invariance as the similar relationships between observed test scores and latent traits across all subgroups.

The data obtained by the measurement instruments are not always complete. For reasons caused by the examinee, the measurement instrument, or the administrator, some data may be missing on data sets.

* This study is based on Mehmet Ali Işıkoğlu's master thesis titled "Comparison of Influence of the Missing Data Handling Methods on Measurement Invarıance".

** Teacher, Ministry of National Education of Republic of Turkey, Ankara-Turkey, mali.isikoglu@gmail.com, ORCID ID: 0000-0001-5104-5661
*** Assoc. Prof. Ph.D., Hacettepe University, Faculty of Education, Ankara-Turkey, burcua@hacettepe.edu.tr, ORCID ID: 0000-0003-3527-686X

To cite this article:
Işıkoğlu, M. A., & Atar, B. (2020). Investigation of the effect of missing data handling methods on measurement invariance of multi-dimensional structures. *Journal of Measurement and Evaluation in Education and Psychology, 11*(3), 312-324. doi: 10.21031/epod.749370.

Received: 08.06.2020
Accepted: 21.09.2020

Missing values arise as a problem since they directly affect the results of the statistical analyses of data sets. As in all other statistical analyses, in the measurement invariance studies, the missing data needs to be checked and managed before the analyses. The presence of missing data can affect the results of many analyses, including confirmatory factor analysis. Since excluding examinees with missing values from data sets will reduce the sample size, the power to generalize the results to the population decreases. In addition, the presence of missing values can cause type I and type II errors. Even the difference in the methods used to handle the missing data problem may lead to different findings from the analysis (Harrington, 2009).

Many techniques have been developed to handle missing data. Allison (2001) classified the missing data handling methods as traditional methods, methods based on Maximum likelihood, and multiple imputation approaches. Listwise deletion (LD) is the method that enables the complete data set to be obtained by removing all cases with unobserved data in any of the variables in the data set. If the missing data has the missing completely at random (MCAR) mechanism, the standard error estimates will be close to the standard error estimates of the real data, since the data set obtained by removing the missing data will be a random sample of the original data set (Allison, 2003). However, if each missing value is in different observations, the sample size will be greatly affected by this situation. This can cause problems even if the missing data has the MCAR mechanism (Enders, 2010). Serial mean imputation (SMI) assigns the mean of the observed data in the variable where the missing data is located, instead of missing data (Little & Rubin, 2002). Since the average of the variable is imputed to the missing data, it does not change the mean value of the variable. However, it reduces the distance of the missing data from the mean to zero, and it underestimates the variance (Enders, 2010; Tabachnick & Fidell, 2013). In the regression imputation (RI) method, the missing variables are imputed values with a regression equation obtained from the observed variables. However, the imputed values have some disadvantages, such as better fit than expected due to estimation from other variable and reducing the variance because it will most likely impute a value close to the mean. And, when the other variables are not a good predictor of the variable with missing value, there is no difference between regression imputation and mean imputation (Tabachnick & Fidell, 2013). Expectation maximization (EM), which is a method based on maximum likelihood, is a method consisting of two steps: expectation (E) and maximization (M), and consists of sequential steps based on a series of regressions. The disadvantage of this method is that the standard errors obtained from this method are not consistent with the actual standard errors (Allison, 2003). In the multiple imputation (MI) method, the random variance is added to the values estimated by regression, unlike EM method. However, different results can be obtained each time due to the addition of random variance (Allison, 2003).

There are two commonly used approaches in measurement invariance tests: confirmatory factor analysis and item response theory (Reise, Widaman & Pugh, 1993). Measurement invariance is generally examined by the multiple-group confirmatory factor analysis (MGCFA) method, which includes hierarchical steps (Whitaker & McKinney, 2007). In order to control the measurement invariance between groups with MGCFA method, configural invariance which requires equality of factor structures between groups, metric invariance which requires equality of factor loadings between groups, scalar invariance which requires equality of intercepts between groups, and strict invariance which requires equality of residual variances between groups must be tested hierarchically (Schoot, Lugtig & Hox, 2012).

Purpose of the Study

The purpose of this study was to investigate the effect of missing data handling methods on measurement invariance of multi-dimensional structures. In this context, the answer to the following problem is sought: "What is the effect of listwise deletion (LD), serial mean imputation (SMI), regression imputation (RI), expectation maximization (EM), and multiple imputation (MI) methods used to handle missing data on the measurement invariance in data sets with different percentages of missingness?".

General Background

In the literature, Reise, Widaman, and Pugh (1993) investigated the effects of confirmatory factor analysis and item response theory models on the invariance of psychological measures. The actual psychological data collected from Minnesota and China were examined by both methods, and their advantages and disadvantages were investigated. Cheung and Rensvold (2002) investigated how GFI goodness of fit statistic changed in MGCFA, which is generally used in measurement invariance studies. As a result of the invariance study performed in the simulation data consisting of two groups, it was suggested to use ΔCFI, ΔGamma, and ΔMcDonald's indices from 20 different fit indices based on GFI. Chen, Wang, and Chen (2012) conducted a simulation study on data sets with different rates of missingness in order to compare the missing data handling methods in exploratory and confirmatory factor analysis. In the study where six different methods were examined, all the methods produced appropriate results for exploratory and confirmatory factor analyses. It was concluded that the most suitable method for exploratory factor analysis was EM. In the case of less than 20% missing, no statistically significant difference was found between the methods. However, when the missing data is more than 30%, it is suggested to use the SMI and linear trend methods It is seen that studies on measurement invariance are generally based on real data among different groups such as gender and culture (Schnabel, Kelava, Vijver & Seifert, 2015; Wang, Willett & Eccles, 2011).Some of the studies were also used to compare the goodness of fit indices used when examining the measurement invariance (Chen, 2007; Cheung & Rensvold, 2002).

Studies on the effect of missing data on test and item parameters and model data fit (Akbaş & Tavşancıl, 2015; Çüm & Gelbal, 2015; Demir, 2013; Köse, 2014) were conducted. However, there are not many studies about the effect of missing data handling methods on measurement invariance under different conditions. In one of these studies, Selvi, Alıcı & Uzun (2020) examined the effect of EM RI, and SMI methods on measurement invariance on the data obtained from the School Attitude Scale developed by Alıcı (2013) under the condition of 5% missing. Findings of the study show that different methods can change measurement invariance decisions. It has been suggested by the researchers to do more research on different missing data structures and different proportions of missing data.

When the studies related to the missing data handling methods were examined, it is generally aimed to determine which method is more successful in handling missing values (Allison, 2003; Chen, Wang & Chen, 2012; Downey & King, 1998; Olinsky, Chen & Harlow, 2003). The data sets used are generally simulation data, and it is seen that the successful methods change in the data sets with different sample sizes and different percentages of missingness. Missing data studies have recently increased. The problem of missing data is no longer ignored, and efforts are being made to solve the problem.

In this context, it is thought that examining the performance of the missing data handling methods at different missing rates in measurement invariance studies on multi-dimensional structures is important in terms of shedding light on the problem of missing data in measurement invariance studies. Five methods frequently used in researches are discussed within the scope of this study.

METHOD

Participants

The sample was 10857 15-years old students (5109 from Turkey and 5748 from Singapore) who participated in PISA 2015 administration from Turkey and Singapore. Students who have fully responded to items on "enjoyment of science, instrumental motivation, and epistemological beliefs about science" were used in the study. Measurement invariance studies between Turkey and Singapore were conducted on a complete data set of 10857 students in total.

Since PISA results are generally used for cross-country comparisons, it was decided to evaluate the measurement invariance between countries in the data set. It was decided to use Turkey and Singapore data whose mean science score distance from the OECD average is approximately equal in absolute value in terms of mean science score. Singapore has 556 mean science score, Turkey has 425 mean science score, and OECD average is 493. It is also taken into account that Singapore is the most successful country in terms of average science score. Similarly, the percentage of variation in science performance explained by students' socio-economic status was also considered.

Data Collection Instruments/Data Collection Methods/Data Collection Techniques

The data used in this study was obtained from the PISA 2015 administration organized by OECD and aimed to evaluate the educational systems of countries. PISA is an administration to measure the level of knowledge and skills necessary for students to participate in modern society. In addition to focusing on key areas such as science, mathematics, and reading, the 2015 administration included collaborative problem solving and financial literacy as an innovative field (OECD, 2016).

In this study, the model including the items of enjoyment of science, instrumental motivation, and epistemological beliefs was used. Enjoyment of science is represented by five items, instrumental motivation by four items, and epistemological beliefs by six items. Each item has four response categories, such as strongly disagree, disagree, agree, and strongly agree. Some sample items are shown in the Table 1.

Table 1. Sample Items of the Model

ST094	How much do you disagree or agree with the statements about yourself below? (Please select one response in each row.)				
		Strongly disagree	Disagree	Agree	Strongly agree
ST094Q01NA	I generally have fun when I am learning <broad science> topics.	$\square_1$	$\square_2$	$\square_3$	$\square_4$
ST094Q02NA	I like reading about <broad science>.	$\square_1$	$\square_2$	$\square_3$	$\square_4$
ST113	How much do you agree with the statements below? (Please select one response in each row.)				
		Strongly disagree	Disagree	Agree	Strongly agree
ST113Q01TA	Making an effort in my <school science> subject(s) is worth it because this will help me in the work I want to do later on.	$\square_1$	$\square_2$	$\square_3$	$\square_4$
ST113Q02TA	What I learn in my <school science> subject(s) is important for me because I need this for what I want to do later on.	$\square_1$	$\square_2$	$\square_3$	$\square_4$
ST131	How much do you disagree or agree with the statements below? (Please select one response in each row.)				
		Strongly disagree	Disagree	Agree	Strongly agree
ST131Q01NA	A good way to know if something is true is to do an experiment.	$\square_1$	$\square_2$	$\square_3$	$\square_4$
ST131Q03NA	Ideas in <broad science> sometimes change.	$\square_1$	$\square_2$	$\square_3$	$\square_4$

Data Analysis

From the complete data set, 5%, 10%, and 20% of values were deleted randomly on the basis of all cells in the dataset with the help of the R program, and missing data with different percentages of missingness were generated. To determine the mechanism of the missing data in the data sets, Little's MCAR test was performed in each data set. MCAR test was examined separately for each country's datasets. Accordingly, for Turkey p= 0.864 (chi-square=3474.455) in the data set with 5% missing, p= 0.909 (chi-square=8279.206) in the data set with 10% missing, and p= 0.921 (chi-square=21341.920) in the data set with 20% missing were found. For Singapore p= 0.976 (chi-square=3458.673) in the data set with

5% missing, p= 0.990 (chi-square=8840.290) in the data set with 10% missing, and p= 0.645 (chi-square=23308.247) in the data set with 20% missing were found. Accordingly, it can be said that the missing data in all data sets have MCAR mechanism. Afterwards, LD, SMI, RI, EM, and MI with five imputation methods were applied to each data set to handle the missing data problem, and inter-country measurement invariance was examined by MGCFA approach on completed data sets.

For cross-country measurement invariance, enjoyment of science, instrumental motivation, and epistemological beliefs model is shown in Figure 1.

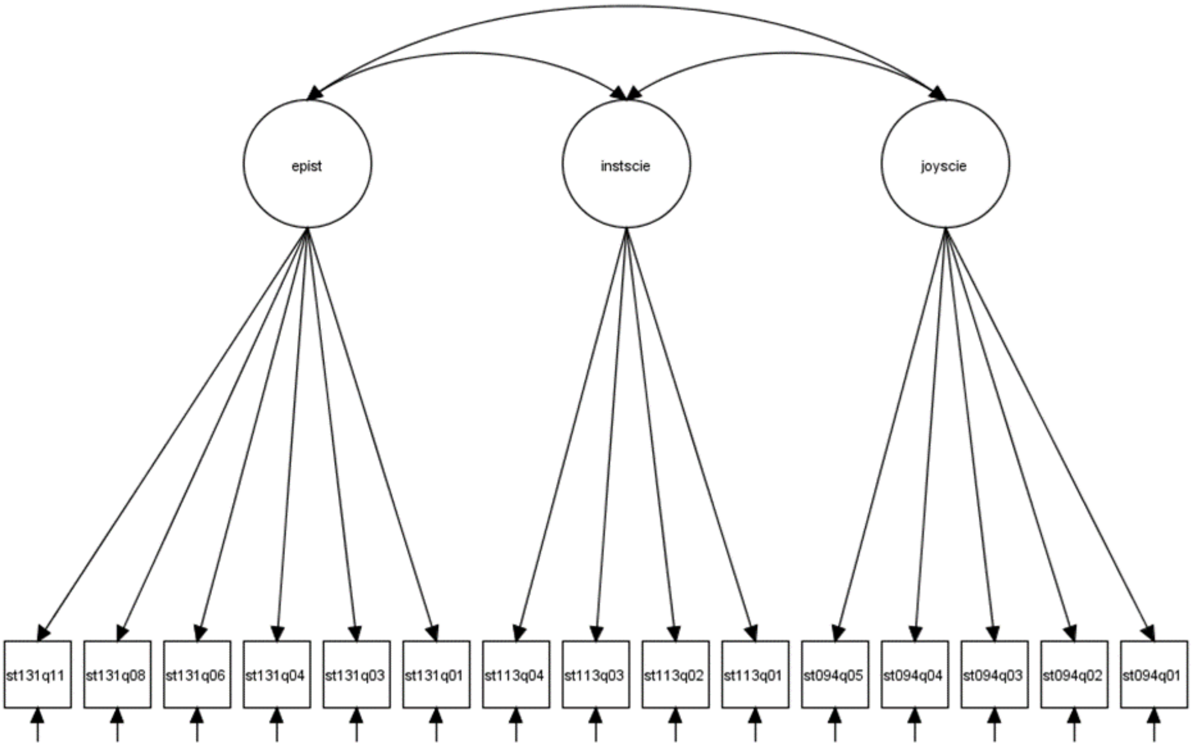

Figure 1. Enjoyment of Science, Instrumental Motivation, and Epistemological Beliefs Model

Before starting the analysis, it is necessary to check the missing values, normality, outliers, and multicollinearity in the data set. The kurtosis and skewness values of each data were examined for normality assumption. According to the findings, the skewness values of the variables ranged from -0.942 to -0.471, and the kurtosis values ranged from -0.296 to 0.913. Tabachnick and Fidell (2013) stated that the closeness of kurtosis and skewness values to zero shows that the distribution is close to normal distribution. According to obtained kurtosis and skewness values, it can be said that each variable was distributed normally. To determine the outliers, z distributions were examined. |z|>3.29 indicates that the variable contains outliers (Tabachnick & Fidell, 2013). According to the findings, z scores of the variables ranged between -2.78 and 1.42. In this case, it can be concluded that there are no outliers in the data set. VIF and tolerance values were examined to determine if there was a multicollinearity problem. VIF values ranged between 2.178 and 4.882, and the tolerance values ranged between 0.205 and 0.459. Based on this finding, it was concluded that there is no multicollinearity problem in the data set.

In order to compare the results obtained from a measurement instrument applied to groups with different characteristics, it is important to ensure the measurement invariance between groups. There are different approaches to test measurement invariance, such as MGCFA and item response theory. In this study, the measurement invariance was examined with the MGCFA approach with ML estimator. MGCFA aims to compare the means, variance, and covariance of the latent variable between the groups while testing the measurement invariance (Asparouhov & Muthen, 2014). In this context, configural invariance, metric invariance, scalar invariance, and strict invariance were tested hierarchically. ΔCFI was examined to determine whether measurement invariance was provided at each stage. A difference of less than .01 supports the less parameterized model (Chung et al., 2016).

RESULTS

In this section, the findings of the research are given. Firstly, the reference values to compare the data sets with different percentages of missingness were obtained by performing a hierarchical measurement invariance in the complete data set.

Before moving on to measurement invariance studies in the whole data set, confirmatory factor analysis was performed in Turkey and Singapore datasets separately, and model-data fits were examined. Fit indices obtained from Turkey and Singapore datasets are presented in Table 2.

Table 2. Fit Indices in the Singapore and Turkey Data Sets

	χ^2	df	χ^2/df	SRMR	RMSEA	CFI	TLI
Singapore	3546.635	87	40.766	.033	.083	.949	.938
Turkey	2036.208	87	23.405	.022	.066	.968	.961

When Table 2 is examined, it is seen that the data for both countries fit the model. After that, a cross-country measurement invariance study was conducted for the complete data set, and reference values were obtained. Reference values obtained from the complete data set are provided in Table 3.

Table 3. Fit Indices in The Complete Data Set

	χ^2	df	χ^2/df	SRMR	RMSEA	CFI	TLI	ΔCFI
Configural	5582.843	174	32.085	.028	.076	.958	.949	
Metric	5723.250	186	30.770	.034	.074	.957	.951	-.001
Scalar	6222.092	198	31.425	.038	.075	.953	.950	-.005
Strict	11469.299	216	53.099	.226	.098	.912	.914	-.046

When the fit indices in the Table 1 were examined, it was seen that configural invariance, metric invariance, and scalar invariance were achieved in the complete data set, but not the strict invariance (|ΔCFI|≤.01). The values related to fit indices from the reference data set was used to compare with the completed data sets. Then, the results of the measurement invariance studies were included in the data sets with 5% missing, 10% missing, and 20% missing and completed with LD, SMI, RI, EM, and MI methods.

Influence of Missing Data Handling Methods on Measurement Invariance in the Data Set with 5% Missing

The data set with 5% missing was completed with LD, SMI, RI, EM, and MI methods, and measurement invariance was hierarchically tested on completed data sets. The fit indices obtained at each stage of measurement invariance according to different methods are provided in Table 4.

Table 4. Fit Indices in the Data Set with 5% Missing and Completed with the Methods

Method	*Invariance*	χ^2	*df*	χ^2/df	*SRMR*	*RMSEA*	*CFI*	*TLI*	*ΔCFI*
LD	Configural	2982.518	174	17.141	.030	.080	.953	.944	
	Metric	3055.602	186	16.428	.035	.078	.952	.946	-.001
	Scalar	3301.304	198	16.673	.039	.079	.948	.945	-.005
	Strict	5661.534	216	26.211	.222	.100	.909	.912	-.044
SMI	Configural	4363.479	174	25.077	.028	.067	.962	.955	
	Metric	4488.792	186	24.133	.033	.065	.961	.956	-.001
	Scalar	4915.130	198	24.824	.037	.066	.958	.955	-.004
	Strict	10062.546	216	46.586	.228	.092	.912	.914	-.050
RI	Configural	5543.385	174	31.856	.028	.075	.958	.949	
	Metric	5661.530	186	30.438	.033	.074	.957	.952	-.001
	Scalar	6139.553	198	31.008	.036	.074	.954	.951	-.004
	Strict	10845.374	216	50.210	.221	.095	.917	.919	-.041
EM	Configural	6153.287	174	35.364	.028	.080	.955	.946	
	Metric	6275.856	186	33.741	.033	.078	.954	.949	-.001
	Scalar	6766.297	198	34.173	.037	.078	.951	.948	-.004
	Strict	11998.191	216	55.547	.228	.100	.912	.914	-.043
MI	Configural	5413.041	174	31.109	.028	.074	.959	.950	
	Metric	5531.746	186	29.741	.033	.073	.958	.952	-.001
	Scalar	6002.515	198	30.316	.036	.073	.954	.951	-.005
	Strict	10786.742	216	49.939	.224	.095	.916	.919	-.040

When the fit indices in the tables were examined, it was seen that the first three stages of measurement invariance between countries were achieved in all data sets, but not the strict invariance (|ΔCFI|≤.01). When the fit indices obtained for each method were compared with the reference values given in Table 1, it was observed that the indices obtained from SMI, RI, EM, and MI methods gave more similar results to the reference values. But dissimilarly, LD and SMI methods showed χ^2/df less than the reference value. All indices, especially ΔCFI, were compared with the reference data set. Methods giving more similar results to the reference values were determined. RI and MI methods yielded the closest results.

Influence of Missing Data Handling Methods on Measurement Invariance in the Data Set with 10% Missing

The data set with 10% missing was completed with LD, SMI, RI, EM, and MI methods. Measurement invariance was hierarchically tested on completed data sets. The fit indices obtained at each stage of measurement invariance according to different methods are provided in Table 5.

Table 5. Fit Indices in the Data Set with 10% Missing and Completed the Methods

Method	Invariance	χ^2	df	χ^2/df	SRMR	RMSEA	CFI	TLI	ΔCFI
LD	Configural	1423.852	174	8.183	.035	.080	.950	.940	
	Metric	1444.182	186	7.764	.038	.078	.950	.944	.000
	Scalar	1526.376	198	7.709	.041	.078	.947	.944	-.003
	Strict	2786.312	216	12.900	.245	.103	.898	.901	-.052
SMI	Configural	3186.870	174	18.315	.025	.056	.969	.963	
	Metric	3275.686	186	17.611	.030	.055	.968	.964	-.001
	Scalar	3625.250	198	18.309	.033	.056	.965	.963	-.004
	Strict	8700.191	216	40.279	.232	.085	.913	.915	-.056
RI	Configural	4943.032	174	28.408	.027	.071	.962	.955	
	Metric	5060.230	186	27.206	.032	.069	.961	.957	-.001
	Scalar	5433.588	198	27.442	.035	.070	.959	.956	-.003
	Strict	9705.969	216	44.935	.217	.090	.925	.927	-.037
EM	Configural	6318.480	174	36.313	.028	.081	.956	.947	
	Metric	6446.346	186	34.658	.033	.079	.955	.949	-.001
	Scalar	6873.543	198	34.715	.036	.079	.952	.949	-.004
	Strict	12000.284	216	55.557	.230	.100	.916	.918	-.040
MI	Configural	4898.776	174	28.154	.027	.071	.962	.954	
	Metric	5015.456	186	26.965	.032	.069	.961	.956	-.001
	Scalar	5407.393	198	27.310	.035	.070	.958	.956	-.004
	Strict	9761.137	216	45.190	.222	.090	.923	.926	-.039

When the fit indices in the tables were examined, it was seen that all the missing data handling methods are provided all the invariance stages except strict invariance as in reference data set ($|\Delta CFI|\leq.01$). When the fit indices obtained for each method were compared with the reference values given in Table 1, it was seen that the EM method gives results very close to the reference values. Dissimilarly, LD and SMI methods showed χ^2/df less than the reference value. And the SMI method showed CFI and TLI values to be more than they were.

Influence of Missing Data Handling Methods on Measurement Invariance in the Data Set with 20% Missing

The data set with 20% missing was completed with LD, SMI, RI, EM,and MI methods, and measurement invariance between countries was hierarchically tested on completed data sets. The fit indices obtained from the measurement invariance studies are provided in Table 6.

Table 6. Fit Indices in the Data Set with 20% Missing and Completed with the Methods

Method	Invariance	χ^2	df	χ^2/df	SRMR	RMSEA	CFI	TLI	ΔCFI
LD	Configural	417.859	174	2.401	.043	.085	.948	.938	
	Metric	425.802	186	2.289	.049	.082	.949	.943	.001
	Scalar	448.435	198	2.265	.051	.081	.947	.944	-.001
	Strict	694.988	216	3.218	.199	.107	.899	.902	-.049
SMI	Configural	2168.515	174	12.463	.023	.046	.973	.968	
	Metric	2227.010	186	11.973	.027	.045	.973	.969	.000
	Scalar	2484.778	198	12.549	.030	.046	.969	.967	-.004
	Strict	7600.786	216	35.189	.239	.079	.901	.904	-.072
RI	Configural	4736.906	174	27.224	.026	.070	.963	.956	
	Metric	4831.875	186	25.978	.030	.068	.963	.958	.000
	Scalar	5153.266	198	26.027	.033	.068	.960	.958	-.003
	Strict	8454.797	216	39.143	.207	.084	.934	.936	-.029
EM	Configural	7818.130	174	44.932	.028	.090	.950	.940	
	Metric	7928.744	186	42.628	.032	.088	.949	.943	-.001
	Scalar	8317.908	198	42.010	.035	.087	.947	.944	-.003
	Strict	13517.876	216	62.583	.234	.107	.913	.916	-.037
MI	Configural	4916.012	174	28.253	.026	.071	.961	.953	
	Metric	4995.423	186	26.857	.030	.069	.960	.955	-.001
	Scalar	5328.987	198	26.914	.033	.069	.958	.955	-.003

	Strict	8937.861	216	41.379	.216	.086	.928	.930	-.033

When the fit indices in the tables were examined, it was seen that all the missing data handling methods provided all the invariance stages except strict invariance (|ΔCFI|≤.01). When the fit indices obtained for each method were compared with the reference values given in Table 1, it was seen that the MI method gives results close to the reference values. The MI method shows χ^2/df close to the reference value, but dissimilarly, LD and SMI methods shows χ^2/df lower than it is, and the EM method shows χ^2/df higher than it is.

DISCUSSION and CONCLUSION

In this study, the effect of completing data sets with missing values with LD, SMI, RI, EM, and MI methods on measurement invariance was investigated. As a result of measurement invariance studies between countries performed in data sets completed with different missing data handling methods in all missing percentages, it was observed that all the invariance stages except strict invariance were provided in accordance with the complete data set. Although the data sets were completed with different methods, there was no result that would show the measurement invariance between countries different from the reference data set.

The research was limited in terms of missing data handling methods, missing data mechanisms, and measurement invariance approaches. LD, SMI, RI, EM, and MI methods were used as missing data handling methods. The data sets have MCAR mechanism. Data sets with multi-dimensional structures were used in the study. And, measurement invariance was handled by MG-CFA approach. The findings and discussion in this study are based on a single data set obtained from the PISA 2015 administration. No replication was done in the study. Please consider this situation as a limitation.

In the literature, methods based on the likelihood approach and the multiple imputation approach are proposed as the strategy of handling the missing data in CFA models (Allison, 2003; Brown, 2006). The findings of the research show that EM and MI methods which are based on the likelihood approach yielded more successful results in accordance with the literature.

Selvi, Alıcı & Uzun (2020) tested the measurement invariance with structural equation modeling in the complete data matrix and in cases of handling the missing data tested using EM, Regression-Based Imputation, and Mean Substitution methods. They concluded that different methods can change the decisions of measurement invariance. But, in the findings of this study it was seen that not all methods change measurement invariance decisions.

Allison (2003) stated that MI has good statistical properties, and it can be used in almost any situation. Schafer and Graham (2002) recommended EM algorithm for maximum likelihood and MI method. Similar to the studies, the results obtained from EM and MI methods were found to be more appropriate to reference data in this study.

As a result of comparing the fit indices obtained from each data set with the fit indices obtained from the complete data set, the data sets completed with RI and MI in the data set with 5% missing yielded closer results to the reference values. In the data set with 10% missing, closer results were obtained from the EM method than the other methods. And in the data set with 20% missing, the missing data handling method which gave the closest results to the reference values was MI. While making comparisons, based on ΔCFI change, the methods whose fit indices give the closest results to the reference values were determined descriptively. As a result of the research, recommendations for implementation are as follows: In the measurement invariance studies to be performed in multi-dimensional data sets, data sets with 5% missing can be completed by RI and MI methods. The EM method works better than other methods if there are around 10% missing. And, if the data set has about 20% missing, the MI method can be used to complete the data set.

REFERENCES

Akbaş, U., & Tavşancıl, E. (2015). Farklı örneklem büyüklüklerinde ve kayıp veri örüntülerinde ölçeklerin psikometrik özelliklerinin kayıp veri baş etme teknikleri ile incelenmesi. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi, 6*(1), 38-57.

Alıcı, D. (2013). Okula yönelik tutum ölçeği'nin geliştirilmesi: Güvenirlik ve geçerlik çalışması. *Eğitim ve Bilim*, 38(268), 318-331.

Allison, P. D. (2001). *Missing data.* Thousand Oaks, CA: Sage.

Allison, P. D. (2003). Missing data techniques for structural equation modeling. *Journal of Abnormal Psychology*, *112*(4), 545-557.

Asparouhov, T., & Muthen, B. (2014). Multiple-group factor analysis alignment. *Structural Equation Modeling: A Multidisciplinary Journal*, *21*(4), 1-14.

Brown, T. A. (2006). *Confirmatory factor analysis for applied research*. New York: The Guilford Press.

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, 14(3), 464-504.

Chen, S. F., Wang, S., & Chen, C. Y. (2012). A simulation study using EFA and CFA programs based the impact of the missing data on test dimensionality. *Expert Systems with Applications*, *39*(4), 4026-4031.

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, *9*(2), 233-255.

Chung, H., Kim, J., Park, R., Bamer, A. M., Bocell, F. D., & Amtmann, D. (2016). Testing the measurement invariance of the University of Washington Self-Efficacy Scale short form across four diagnostic subgroups. *Qual Life Res.*, *25*(10), 2559-2564.

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

Demir, E. (2013). Kayıp verilerin varlığında çoktan seçmeli testlerde madde ve test parametrelerinin kestirilmesi: SBS örneği. *Eğitim Bilimleri Araştırmaları Dergisi, 3*(2), 47-68.

Downey, R. G. & King, C. V. (1998). Missing data in likert ratings: A comparison of replacement methods. *The Journal of General Psychology*, *125*(2), 175-191.

Drasgow, F. (1984). Scrutinizing psychological tests: Measurement equivalence and equivalent relations with external variables are the central issues. *Psychological Bulletin*, *95*(1), 134-135.

Enders, C. K. (2010). *Applied missing data analysis.* New York: The Guilford Press.

Harrington, D. (2009). *Confirmatory factor analysis*. New York: Oxford University Press.

Köse, A. (2014). The effect of missing data handling methods on goodness of fit indices in confirmatory factor analysis. *Educational Research and Reviews, 9*(8), 208-215.

Little, R. J. A. & Rubin, D. B. (2002). *Statistical analysis with missing data. (2nd edition)*. New York: Wiley

OECD (Organization for Economic Cooperation and Development) (2016). *PISA 2015 results in focus*. Retrieved February 12, 2017, from https://www.oecd.org/pisa/pisa-2015-results-in-focus.pdf

Olinsky, A., Chen, S., & Harlow, L. (2003). The comparative efficacy of imputation methods for missing data in structural equation modeling. *European Journal of Operational Research*, *151*(1), 53-79.

Reise, S. P., Widaman, K. F., & Pugh, R. H. (1993). Confirmatory factor analysis and item response theory: Two approaches for exploring measurement invariance. *Psychological Bulletin*, *114*(3), 552-566.

Schafer, J. L. & Graham, J. W. (2002) Missing data: Our view of the state of the art. *Psychological Methods*, *7*(2), 147-177.

Schnabel, D. B. L., Kelava, A., Vijver, F. J. R., & Seifert, L. (2015). Examining psychometric properties, measurement invariance, and construct validity of a short version of the Test to Measure Intercultural Competence (TMIC-S) in Germany and Brazil. *International Journal of Intercultural Relations, 49*, 137-155.

Schoot, R., Lugtig, P., & Hox, J. (2012). A checklist for testing measurement invariance. *European Journal of Developmental Psychology,* 9(4), 486-492.

Selvi, H., Alıcı, D., & Uzun, N. B. (2020). Investigating measurement invariance under different missing value reduction methods. *Asian Journal of Education and Training, 6*(2), 237-245.

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

Wang, M., Willett, J. B., & Eccles, J. S. (2011). The assessment of school engagement: Examining dimensionality and measurement invariance by gender and race/ethnicity. *Journal of School Psychology*, *49*(4), 465-480.

Whitaker, B. G., & Mckinney, J. L. (2007). Assessing the measurement invariance of latent job satisfaction ratings across survey administration modes for respondent subgroups: A MIMIC modeling approach. *Behavior Research Methods, 39*(3), 502-509.

Xu, H., & Tracey, T. J. G. (2017). Use of multi-group confirmatory factor analysis in examining measurenet invariance in counseling psychology research. *The European Journal of Counselling Psychology*, *6*(1), 75-82.

Çok Boyutlu Yapılarda Kayıp Veri ile Baş Etme Yöntemlerinin Ölçme Değişmezliğine Etkisi Açısından Karşılaştırılması

Giriş

Ölçme araçları, eğitim sistemleri içerisinde büyük öneme sahiptir. İhtiyaca uygun nitelikli insan gücü yetiştirmek için bireylerin eğitim kurumlarına yerleştirilmesi, eğitim sistemlerinde ihtiyaca uygun olarak değişiklikler ve geliştirmeler yapılabilmesi ölçme araçlarından elde edilen bulgular sonucunda yapılabilmektedir. Ulusal ve uluslararası düzeyde yapılan ölçme ve değerlendirme çalışmaları sonucunda, ülkeler eğitim politikalarında dahi değişikliklere gidebilmektedir. Özellikle uluslararası karşılaştırmaların yapılmasına olanak sağlayan büyük ölçekli ölçme ve değerlendirme çalışmalarının sonuçları, eğitimin tüm paydaşları tarafından ilgiyle takip edilmektedir.

Uluslararası düzeyde uygulanan PISA (Programme for International Student Assessment) ve TIMSS (Trends in International Mathematics and Science Study) projeleri, ülkeler arası karşılaştırma yapılmasını amaçlamaktadır. Özellikle uluslararası düzeyde yapılan ve sonucunda karşılaştırma yapılan sınavlarda, elde edilen sonuçların karşılaştırılabilir olması, ülkeler arası değerlendirmelerde büyük önem taşımaktadır. Aynı ölçme aracının uygulandığı, özellikleri birbirinden farklı gruplardan elde edilen bulguların yorumlanabilmesi için, ölçme aracının bütün gruplar için aynı anlama gelmesi gerekmektedir. Bu bağlamda karşımıza ölçme değişmezliği kavramı ortaya çıkmaktadır. Drasgow (1984) ölçme değişmezliğini "gözlenen test puanları ile gizil özelliklerin arasındaki ilişkinin tüm alt gruplar arasında benzer olması" şeklinde tanımlamıştır.

Ölçme araçları tarafından elde edilen veriler her zaman eksiksiz şekilde elde edilememektedir. Yanıtlayıcıdan, ölçme aracından veya uygulayıcıdan kaynaklanan sebeplerden dolayı veri setlerinde, bazı değişkenlerde kayıp veriler bulunabilmektedir. Kayıp veriler, veri setleri üzerinde yapılan istatistiksel işlemlerin sonuçlarını doğrudan etkileyen önemli bir problemdir.

Bu araştırmanın amacı, kayıp veri ile baş etme yöntemlerinin ölçme değişmezliğine etkisi açısından karşılaştırılmasıdır. Yapılan ölçme değişmezliği çalışmalarında kayıp veri durumu, analizlere başlamadan önce kontrol edilmesi ve çözülmesi gereken bir problemdir. Kayıp verilerin varlığı, doğrulayıcı faktör analizi de dahil olmak üzere birçok analizin sonucunu etkileyebilir. Kayıp verilerin veri setinden çıkarılması örneklemi küçülteceğinden, elde edilen sonuçların evrene genellenebilme gücü azalır. Ayrıca kayıp verilerin varlığı, analizlerden elde edilen anlamlılık değerlerini etkileyerek tip I ve tip II hataların oluşmasına sebep olabilir. Kayıp veri problemini çözmek için kullanılan yöntemlerin farklılığı bile, analizlerden farklı bulgular elde edilmesine neden olabilir (Harrington, 2009).

Kültür, etnik köken, dil, cinsiyet gibi farklı gruplardan bireylerin karşılaştırılmasında kullanılan testlerin öncelikle ölçme değişmezliğinin sağlanması gerekmektedir. Ölçme değişmezliği analizlerinden elde edilen bulguların doğru bir şekilde yorumlanabilmesi için ise kayıp veri probleminin çözülmesi gerekmektedir. Farklı kayıp veri oranlarına bağlı olarak, kayıp veri ile baş etme yöntemlerinin karşılaştırıldığı bu araştırmadan elde edilen bulgularla, yapılacak olan ölçme değişmezliği çalışmalarında veri setine uygun olan kayıp veri ile baş etme yönteminin seçilebilmesi amaçlanmıştır. Bu bağlamda "Kayıp veriler ile baş etmede kullanılan dizin silme (DS), seri ortalaması atama (SO), regresyon atama (RA), beklenti maksimizasyonu (BM) ve çoklu atama (ÇA) yöntemlerinin, farklı oranlarda kayıp içeren veri setlerinde ölçme değişmezliğine etkisi ne düzeydedir?" problemine yanıt aranmaktadır.

Alan yazın incelendiğinde, Reise, Widaman ve Pugh (1993), doğrulayıcı faktör analizi ve madde tepki kuramı modellerinin, psikolojik ölçmelerin değişmezliğine etkilerini araştırmışlardır. Minesota ve Çin'den toplanan gerçek psikolojik veriler her iki yöntemle de incelenmiş ve yöntemlerin avantaj ve dezavantajları araştırılmıştır. Cheung ve Rensvold (2002), ölçe değişmezliği çalışmalarında genellikle kullanılan çok gruplu doğrulayıcı faktör analizinde, GFI uyum iyiliği istatistiğinin ne şekilde değiştiğini araştırmışlardır. İki gruptan oluşan simülasyon verisinde gerçekleştirilen değişmezlik çalışmasının sonucunda, GFI indeksini temel alan 20 farklı uyum indeksinden, ΔCFI, ΔGamma ve ΔMcDonald's indekslerinin kullanılması önerilmiştir. Chen, Wang ve Chen (2012), açımlayıcı ve doğrulayıcı faktör analizinde kayıp veri yöntemlerini karşılaştırmak amacıyla, farklı oranlarda kayıp içeren veri setleri üzerinde simülasyon çalışması yapmıştır. Altı farklı yöntemin incelendiği çalışmada, tüm yöntemler açımlayıcı ve doğrulayıcı faktör analizinde modele uygun sonuçlar üretmiştir. Açımlayıcı faktör analizi için en uygun yöntemin beklenti maksimizasyonu olduğu sonucuna ulaşılmıştır. %20'nin altında kayıp olması durumunda yöntemler arasında istatistiksel olarak anlamlı bir fark bulunamamıştır. Ancak eksik veriler %30'dan fazla olduğunda, seri ortalaması atama yöntemi ve doğrusal eğim yöntemi kullanılması önerilmiştir.

Dünyada ve Türkiye'de ölçme değişmezliği ile ilgili yapılan çalışmalara genel olarak bakıldığında, çalışmaların genelinde gerçek veriler kullanılarak cinsiyet ve kültür gibi farklı gruplar arasında ölçme değişmezliğinin sağlanıp sağlanmadığıyla ilgili olduğu görülmektedir (Schnabel, Kelava, Vijver ve Seifert, 2015; Wang, Willett ve Eccles, 2011;). Bir kısım araştırmaların da, ölçme değişmezliği incelenirken kullanılan uyum iyiliği katsayılarının karşılaştırılmasına yönelik olduğu görülmüştür (Chen, 2007; Cheung ve Rensvold, 2002;).

Kayıp veri atama yöntemleri ile ilgili olan çalışmalara bakıldığında ise, çalışmalar genellikle kayıp verilerin tamamlanmasında hangi yöntemin daha başarılı olduğunu belirlemeye yöneliktir (Allison, 2003; Chen, Wang & Chen, 2012; Downey & King, 1998; Olinsky, Chen & Harlow, 2003). Kullanılan veri setleri genellikle simülasyon verileri olup, başarılı yöntemlerin, farklı örneklem büyüklüklerinde ve farklı oranlarda kayıp içeren veri setlerinde değiştiği görülmektedir. Kayıp veri çalışmaları son zamanlarda artmıştır. Kayıp veri sorunu, artık göz ardı edilmeyerek, problemin çözümüne yönelik çalışmalar yapılmaktadır. Bu araştırmada, kayıp veri atama yöntemlerinin ölçme değişmezliğine etkisi araştırılmaktadır. Alan yazında yapılacak ölçme değişmezliği çalışmalarında, farklı örneklem büyüklüklerinde ve farklı oranlardaki kayıp verilerde, kayıp veri probleminin çözümüne yönelik öneriler getirmek amaçlanmaktadır.

Yöntem

Bu araştırmada, farklı kayıp veri ile baş etme yöntemleri ile tamamlanmış veri setlerinde ölçme değişmezliği çalışması yapılmıştır. Çalışmanın amacı, DS, SO, RA, BM ve ÇA yöntemlerinin çok boyutlu yapılarda ölçme değişmezliğine etkisini incelemektir.

Araştırmanın örneklemini PISA 2015 uygulamasına Türkiye ve Singapur'dan katılmış 12010 (Türkiye=5895, Singapur=6115) öğrenciden, fen okuryazarlığına ilişkin duyuşsal özellikler ile ilgili maddelere eksiksiz yanıt vermiş 10857 (Türkiye=5109, Singapur=5748) öğrenci oluşturmaktadır.

Araştırmada, 5496 kişilik eksiksiz veri setinde ülkeler arası ölçme değişmezliği çalışmaları yapılmıştır.

Oluşturulan eksiksiz veri setinden, hücre bazında %5, %10 ve %20 oranında rastgele değerler R programı yardımıyla silinmiş ve kayıp veriler oluşturulmuştur. Oluşturulan veri setlerinde bulunan kayıp verilerin mekanizmasının belirlenebilmesi için, her veri setinde Little'ın TROK testi gerçekleştirilmiştir. Türkiye örneklemi için %5 kayıp içeren veri setinde p= 0,864 (ki-kare=3474,455), %10 kayıp içeren veri setinde p= 0,909 (ki-kare=8279,206) ve %20 kayıp içeren veri setinde p= 0, 921 (ki-kare=21341,920) bulunmuştur. Singapur için %5 kayıp içeren veri setinde p= 0, 976 (ki-kare=3458,673), %10 kayıp içeren veri setinde p= 0, 990 (ki-kare=8840,290) ve %20 kayıp içeren veri setinde p= 0, 645 (ki-kare=23308,247) bulunmuştur. Buna göre tüm veri setindeki kayıp verilerin TROK mekanizmasına sahip olduğu söylenebilir.

Daha sonra, her bir veri setinde DS, SO, RA, BM ve ÇA yöntemleri uygulanmış ve ülkeler arası ölçme değişmezliği çalışmaları çok gruplu doğrulayıcı faktör analizi (ÇGDFA) yaklaşımı ile incelenmiştir.

Sonuç ve Tartışma

Araştırmada kayıp veri içeren veri setlerinin, DS, SO, RA, BM ve ÇA yöntemleriyle tamamlanmasının, ölçme değişmezliğine etkisi araştırılmıştır. Tüm oranlarda, farklı yöntemlerle tamamlanmış veri setlerinde yapılan ülkeler arası ölçme değişmezliği çalışmalarının sonucunda, eksiksiz veri setine uygun olarak katı değişmezlik dışındaki tüm değişmezlik aşamalarının sağlandığı görülmüştür. Veri setlerinde, farklı kayıp veri ile baş etme yöntemleri ile tamamlansa da, ülkeler arası ölçme değişmezliğini referans veri setinden farklı gösterecek bir sonuç bulunmamıştır.

Araştırma, kayıp veri ile baş etme yöntemleri, kayıp veri mekanizması ve ölçme değişmezliği yaklaşımı açısından sınırlandırılmıştır. Kayıp veri ile baş etme yöntemlerinden dizin silme (DS), seri ortalaması atama (SO), regresyon atama (RA), beklenti maksimizasyonu (BM) ve çoklu atama (ÇA) yöntemlerine yer verilmiştir. Veri setleri, kayıp veri mekanizmalarından tamamen rassal olarak kayıp (TROK) mekanizmasına sahiptir. Çalışmada birden fazla faktöre sahip veri setleri kullanılmıştır. Ölçme değişmezliği çok gruplu doğrulayıcı faktör analizi yaklaşımıyla ele alınmıştır.

Her bir veri setinden elde edilen uyum katsayılarının, eksiksiz veriden elde edilen uyum katsayıları ile karşılaştırılması sonucunda, %5 kayıp içeren veri setinde RA ve ÇA ile tamamlanan veri setleri, referans değerlere daha yakın sonuçlar vermiştir. %10 kayıp içeren veri setinde, BM yönteminden diğer yöntemlere göre daha yakın sonuçlar elde edilmiştir. %20 kayıp içeren veri setinde ise referans değere en yakın sonuç veren kayıp veri ile baş etme yöntemi ÇA olmuştur.

Araştırma sonucunda uygulamaya yönelik olarak öneriler şu şekildedir:

Çok boyutlu veri setlerinde yapılacak olan ölçme değişmezliği çalışmalarında, %5 civarında kayıp içeren veri setleri RA veya ÇA yöntemi ile tamamlanabilir. %10 civarında kayıp veri bulunuyorsa BM yöntemi diğer yöntemlere göre daha iyi sonuç vermektedir. Kayıp veri miktarı %20 civarında ise, ÇA yöntemi kayıp verileri tamamlamak için kullanılabilir.