International Online Journal of Educational Sciences, 2017, 9 (3), 603 - 616

iojes
www.iojes.net

International Online Journal of Educational Sciences

ISSN: 1309-2707

Examination of the Effects of Different Missing Data Techniques on Item Parameters Obtained by CTT and IRT

Ufuk Akbaş[1]

[1]*Hasan Kalyoncu University, Department of Measurement and Evaluation in Education, Turkey*

ARTICLE INFO

Article History:
Received 10.04.2017
Received in revised form 11.05.2017
Accepted 04.06.2017
Available online 24.07.2017

ABSTRACT

The purpose of this study is to examine the effect of different missing data techniques on the item parameters estimated for Classical Test Theory (CTT) and Item Response Theory (IRT) comparatively through simulated and real data sets. For this purpose, data sets with missing completely at random pattern with different sample sizes and missing data rates have been generated. Item parameters based on CTT and IRT are estimated after applying different missing data techniques (listwise deletion, regression imputation and expectation – maximization). Estimated parameters were compared with the parameters of complete data sets. The same procedure was performed on the data of the PM3 subtest in the PISA 2012 application. It is found that higher difficulty and lower discrimination values are obtained with listwise deletion while accurate item difficulties can be obtained with regression imputation and expectation - maximization algorithm, but the same situation is not valid for item discriminations. It is also seen that missing data at a ratio of %2 can lead to serious problems even if they have missing completely at random pattern.

Keywords:
Missing data, missing data techniques, classical test theory, item response theory, item parameters

Introduction

Missing data is a common problem confronted in almost every survey. Empty cells in the data collected within the scope of the research may cause findings and interpretations to be biased and may reduce the effectiveness of the data collection process. There are different disclosures about missing data and its possible causes in the literature. Unit nonresponse is the case when all data is missing for an individual, and item nonresponse is the case when there are unobserved cells within any variable (Heerwegh, 2005). Field (2005) states that missing data may be seen if some of the items are mistakenly skipped, mechanical measuring instruments are malfunctioning, or in the case of carelessness at data entry. Schafer and Graham (2002) define missing data as wave or dropout for especially longitudinal studies. Missing data is considered as wave, when participants may be present for some waves of data collection and not for others; and as dropout, when one leaves the study and does not return.

The commonly accepted classification for missing data is attributed to the works of Rubin (1976) and Little and Rubin (1987). According to this classification, any set of data which has missing observations can have missing completely at random, missing at random, and missing not at random patterns. Shortly, missing completely at random (MCAR) pattern means that the probability of being missing of any data is unrelated to other variables measured within the context of the study, and the relevant variable is not related to its value. Missing at random (MAR) pattern refers to cases where this probability is related to another variable. Missing not at random (MNAR) pattern is the case when the probability of data to be missing is related to its own value (Enders, 2010).

[1] Corresponding author's address: Hasan Kalyoncu University, Department of Measurement and Evaluation in Education, Gaziantep, Turkey
Telephone: +903422118080
Fax: +903422118081
e-mail: ufuk.akbas@hku.edu.tr
DOI: https://doi.org/10.15345/iojes.2017.03.002

The seriousness of the problem caused by missing data depends on the pattern and the amount; and the pattern is a more serious problem than the amount of missing data. There is also no clear criterion as to what amount of missing data can be tolerated at which sample size (Tabachnick and Fidell, 1996). While there is an agreement in the literature that small amount of missing data in large data sets does not cause a significant problem (Tabachnick and Fidell, 1996; Field, 2005), the ratios specified for the amount of ignorable missing data, which have MCAR pattern, have been found inconsistent at %5 (Schafer; 1997; SPSS, 2007) and %10 (Hair, Black, Babin, Anderson and Tatham, 2006).

Leeuw, Hox, and Huisman (2003) indicate that data collection processes and tools, the characteristics of the individuals involved in the sample and even the data entry process should be carefully examined in the presence of missing observations in the data set. From this point of view, it can be said that it is necessary but inadequate to determine the amount of missing data before the analysis of the data obtained in a research starts. Along with the quantity, it is necessary to examine whether the missing data exhibit a certain pattern, in other words, randomness. The pattern of missing data can be examined in several different ways.

Tabachnick and Fidell (1996) indicate that the variables in the dataset are divided into two groups with missing data (0) and complete data (1), and the significance of the difference between means over other variables can be examined by considering these groups. Non-significant results give the information that the distribution of data has a MCAR pattern. Similar to this, Alpar (2011) suggests coding variables as 0-1 and examining the correlation between variable pairs. The low correlation coefficients indicate randomness. If this randomness is the case for all variable pairs, the researcher may conclude that missing data has MCAR pattern.

Another inquiry for randomness can be examined with Little's MCAR test developed by Little (1988). The test is based on the grouping of the rows with the same missing data pattern in the data set and the comparison of their means (Enders, 2010). This test is available in the missing data module found in the latest versions of the widely-used software, SPSS.

In addition, descriptive investigations, such as whether missing data tends or not to occur in any single variable, determining which combinations of variables have unobserved points and how often they are encountered provide important information about randomness. At this point, it is important for researchers to examine the randomness with multiple methods.

Since statistical analyzes to be performed after these examinations require the data to be complete, it is not possible not to use any missing data techniques. This situation has provided the basis for the emergence of different missing data techniques during the data analysis phase. When the literature is examined, it is seen that there are a lot of missing data techniques. While some of these techniques are based on removing cases, others impute single value and model based techniques such as multiple imputation results with more than one possible imputed data sets. Here, frequently used techniques are summarized regarding to their basic characteristics and similarities / differences.

Listwise deletion: In this technique, only rows which have full observations are used. In other words, rows that have at least one missing value are removed. In order to be able to use this technique, which is the most simple and direct approach to the solution of the problem of missing data in the majority of statistical software, data is necessary to have MCAR pattern (Alpar, 2011). It is also seen that this technique is usually used to obtain a lower limit for estimations in comparative studies of missing data techniques (Ginkel, Ark and Sijtsma, 2007; Akbaş, 2014). In addition, it should be noted that this technique has a negative effect on sample size.

Pairwise deletion: The basic feature of this technique is that all the available data is used to make calculations (Allison, 2002). For example, when the covariance needs to be calculated for the variables x and y, the rows with complete values in the variables x and y are considered. For covariance between the x and z variables for the same data set, the calculations are performed over pairs that have the complete data for x and z variables. So, estimations may rely on different subsamples for the same data set. Roth (1994) notes that listwise deletion and pairwise deletion techniques are the most frequently used techniques to deal with missing data.

Mean imputation: The mean calculated over complete part of any variable is imputed to the empty cells in this variable. The basis of this technique is that the mean is the best point estimate for a variable and there are also applications such as the imputation of the group mean according to a group which made up from a variable that is different from the variable to be imputed (Hair et al., 2006). With an approach similar to the mean imputation, for the missing data in categorical variables, mode (Buuren, 2012), and for the skewed distributions, the median (McKnight, McKnight, Sidani and Figueredo, 2007) can be imputed.

Regression imputation: With this technique, missing data is estimated over other variables that do not contain missing data. The regression equation in which the variable to be imputed is considered as the dependent and the other variables as the predictor variable is established. The data is completed by adding the known values of the predictor variables in equation (Enders, 2010).

Expectation - Maximization Algorithm: The values to be imputed are determined by the iteration of the maximization step that the regression equation established over the completed data set and the expectation step that the estimations are obtained by rearranging regression equation (Enders, 2010). Schafer and Graham (2002) indicate that using expectation – maximization (EM) algorithm under MCAR pattern, unbiased parameter estimates can be achieved.

Multiple imputation: When this technique is used, each missing value in the data set is imputed with different possible values of m>1 number. These values are determined based on a distribution determined for each missing data. While there is no change in the exact parts of the original data set, each set of data is different when considered together with the imputed values. In the next steps, each data set is analyzed separately and the results obtained are combined (Buuren, 2012). Combining parameters is based on averaging different estimations. And for standard errors, within imputation variance and between imputation variance is calculated separately (Enders, 2010).

Techniques such as linear interpolation, linear trend at point, imputation of the mean / median of the near points are influenced by the row or column of the data set of imputed data. It would be appropriate to use techniques such as last observation carried forward and next observation carried backward in longitudinal studies where repeated measurements are observed (Akbaş, 2014).

Allison (2002) states that some techniques perform better than others, but there is no technique to deal with a certain level of missing data which can be used in all situations and which can be claimed to be "the best". This statement explains that missing data and many techniques to handle with missing data are used within the scope of many researches.

There are many studies which investigate, the reliability and validity of missing data techniques (Enders, 2003, 2004, Bernaards and Sijtsma, 1999, Çokluk and Kayri, 2011, Akbaş 2014, Nartgün, 2015), t test and ANOVA parameters (Köse and Öztemur, 2014), SEM model - data adaptation (Çüm and Gelbal, 2015) within the context of descriptive statistics (Bal, 2003, Şahin Kürşad, 2014). In addition to these, there are also studies in which Cronbach α and Mokken scalability coefficients (Sijtsma and van der Ark, 2003) and mean squared errors (Holman and Glas, 2005) are examined within the scope of IRT.

Finch (2008) examined seven different missing data techniques through 3-PL models. The study included 20 sample data sets for different sample size, missing data rate and missing data patterns. On item level investigations for the items which were relatively low in difficulty and discrimination (b_1=-.33, a_1=.44 and b_2=-2.7, a_2=.76) with EM, discrimination was found to be positive; for the items which were relatively high in difficulty and discrimination (b_3=1.28, a_3=1.02 and b_4=.57, a_4=1.32) discrimination was found to be negative biased. The item difficulties estimated by EM were found to be negative in all conditions.

Demir (2013) examined the effectiveness of 12 different missing data techniques on the SBS 2011 math test booklet data (n = 527517). It has been determined that there are %2 to %34.4 missing data rates for 20 items in the mathematics test and that the assumption of MCAR is not met. Since the situation in which the data is missing is not known, the performances of the missing data techniques have been examined comparatively through EFA, CFA, item parameters, test parameters and internal consistency coefficients. It is seen that the item difficulties and discriminations obtained by the regression imputation and EM are similar to each other, but lower than those obtained by listwise deletion.

Doğanay Erdoğan (2012) compared the multiple imputation techniques on the person and item parameters obtained with Rasch models. In the first stage of the research, the analyzes carried out on simulated data sets on the basis of PCM and the same procedure was repeated on a real data set in the second stage. When the performances of missing data techniques are examined comparatively through the graphs, it is seen that when the missing data rate is %10 under MCAR pattern, the estimated b parameters are largely similar to those obtained from complete data, when the missing data rate is %30 and %50, significant differences emerge. The findings obtained from the simulated and real data sets included in the study are found to be similar to each other.

When the current literature is examined, it is seen that item parameters of missing data techniques estimated within the scope of CTT and IRT need to be compared. Findings to be reached by this research are expected to lead the researcher in case of encountering missing data in studies such as large scale tests, test development, equating, and item banking studies which will be carried out for both theories.

Method

Design of the Research

The research is in the nature of a basic research in which the effect of different missing data techniques on item parameters estimated on the basis of CTT and IRT is investigated comparatively with simulated and real data sets. Basic researches are researches aiming to add more on the existing knowledge (Karasar, 2007). In the research, missing data techniques are given with an explanatory approach.

This research consists of two main and one additional studies. In the first study, the effectiveness of three different missing data techniques was compared over simulated data sets; in the second study, the same comparison was made over the complete PISA 2012 PM3 subtest data. In the additional study, a new data set with the same sample size as the PM3 subtest was produced and compared with the results obtained in the first two studies.

When similar studies (Bal, 2003; Sijstma and van der Ark, 2003; Holman and Glas, 2005; Finch, 2008; Doğanay Erdoğan, 2012; Şahin Kürşad, 2014) are examined, it can be seen that the sample size is generally manipulated between 50 – 2000, and the missing data rate changed between %1 and %50. It has been observed that the data sets included in these studies have items in the range of 10 - 30. Çakıcı Eser (2015) suggests that it is appropriate to include 12 items in one-dimensional tests so that item parameter estimates can be improved, and the PISA 2012 PM3 subtest contains 12 items. The rationale for selection of PISA 2012 data is that this data is well studied. With reference to these studies, the data sets are set as 500, 1000 and 2000 to represent small, medium and large samples; missing data rate is %2, %5 and %10, respectively, having MCAR pattern and the number of items in the data set is fixed to be k = 12. It has been noted that the rate of $\chi 2/sd$ is less than 3 for the evaluation of model fit within the scope of IRT. By emphasizing 2 parameter model, it is aimed to compare the item discrimination and item difficulty indices according to the CTT and IRT. In the study, listwise deletion is included since it is often and by default used; regression imputation technique is included since it is considered simple and EM is included since it is accepted as an advanced missing data technique.

Data Sets

In the first stage of the study, the effects of three different types of missing data techniques under different conditions were examined through simulated data sets. Hambleton, Swaminathan and Rogers (1991) indicate that difficulty parameters usually vary between -2.0 and 2.0 and item discrimination parameters usually vary between .00 and 2.00 and when related studies (Can, 2003; Çelen, 2008; Özer Özkan, 2012) are investigated it can be seen that these parameters mostly vary between a narrower range. The sample sizes are 500, 1000 and 2000; data sets with item difficulties [-1, +1] and discriminations [0.5 - 1.5] were produced with WinGen 2 (Han, 2007) software. Data sets were generated by performing thirty replications, with the item parameters for the sample sizes included in the study, at specified intervals. Table 1 shows the mean values of the item parameters (p: difficulty and d: discrimination) obtained according to CTT and IRT (b: difficulty and a: discrimination) over the data sets generated and the χ^2/sd rates for model fit.

Table 1. Means of item parameters and χ2/sd rates obtained according to CTT and IRT from simulated data sets

Item no	n=500		n=1000		n=2000		n=500 (χ2/sd=1,16)			n=1000 (χ2/sd=1,82)			n=2000 (χ2/sd=2,67)		
	p	d	P	d	p	d	b	a	χ2/sd	b	a	χ2/sd	b	a	χ2/sd
I1	.71	.82	.31	.60	.22	.78	-.66	1.61	.83	.76	.90	1.89	.99	1.36	2.51
I2	.64	.64	.38	.67	.45	.83	-.52	.92	.89	.43	1.09	2.13	.13	1.65	2.54
I3	.43	.58	.44	.45	.29	.74	.28	.82	1.10	.29	.56	1.55	.73	1.17	2.68
I4	.41	.58	.80	.75	.69	.74	.37	.84	1.30	-.98	1.74	2.50	-.60	1.66	2.65
I5	.72	.53	.37	.53	.21	.86	-1.03	.71	1.09	.56	.71	1.93	.94	1.83	2.73
I6	.38	.48	.73	.63	.76	.70	.55	.65	.95	-.89	.98	1.67	-.85	1.54	2.62
I7	.68	.85	.24	.60	.40	.60	-.55	1.88	.90	1.03	.96	2.03	.42	.79	2.72
I8	.33	.61	.69	.64	.32	.80	.63	1.00	2.30	-.69	1.01	1.38	.58	1.41	2.67
I9	.60	.83	.63	.47	.69	.55	-.30	1.72	1.20	-.65	.59	1.48	-.79	.77	2.28
I10	.53	.50	.43	.69	.56	.54	-.11	.66	1.07	.25	1.14	1.96	-.27	.69	2.64
I11	.48	.50	.53	.68	.38	.87	.08	.65	1.08	-.10	1.11	1.68	.36	1.88	2.60
I12	.77	.68	.30	.47	.20	.82	-1.05	1.06	1.24	.97	.62	1.68	1.02	1.62	2.70

When Table 1 is examined, it is seen that for samples of sizes 500, 1000 and 2000, the range of b parameters obtained on the basis of IRT are [-1.05-.63], [-.98-1.03] and [-.79-.99] respectively, and a parameters change in the range [.65-1.72], [.56-1.74] and [.79-1.88]. It has been recognized that item parameters represent negligible deviations from the values established during the data production phase. It is seen that the rates of χ^2/sd, which indicates data-model and item-model fit, are smaller than 3 for all conditions.

The cognitive data sets of the PISA 2012 achievement test, which was examined during the second phase of the this study, are available at https://www.oecd.org/pisa/pisaproducts/pisa2012database-downloadabledata.htm. From this file containing data from all the countries participating in the PISA 2012 study, only the data belonging to the Turkish sample were selected. In the study, a total of 13 question sets from seven mathematics, three reading and three science fields were included in each of four different booklets (OECD, 2014). Before the real data set included in the second part of the study was determined, the distribution of questionnaires to the booklets, the number of questions they contained, whether there were any items partially scored and the number of rows which had missing data were considered. As a result of these examinations, PM3 mathematics subtest, which included 12 items scored as 1-0, and 1482 students answered in a complete way, seemed appropriate.

Process

A special software has been developed to allow deletion of %2, %5 and %10 rates on the complete data sets to make it have MCAR pattern. This software matches each cell in the data set starting from one to different values. In other words, cell one is in the 1st column of the 1st row and last cell (kxn) is in the last column (k) of the last row (n). The developed software selects random numbers with a lower bound of 1, an upper bound of nxk, and the data in the corresponding cell is deleted. If a data in any cell has already been deleted, a new random value is selected. This cycle continues until the requested rate of missing data is reached in the data set.

With this software developed, %2, %5 and %10 data deletion was performed on data sets with different sample sizes. After this process, it was examined by Little's MCAR test whether the MCAR condition was satisfied and it was seen that the test results for all data sets were not significant (p>.05). In other words, the distribution of the cells with missing data in the dataset was not different from the random distribution.

Imputation of values with regression and EM on data sets containing different rates of missing data were performed using the missing data module in SPSS 22. Necessary commands added to BILOG-MG syntax files so that IRT parameter estimates can be reached since there is no value imputation with listwise deletion. On the complete data set of the PM3 subtest, regression imputation and EM processes were performed after data deletion process at %2, %5, and %10 rate which had MCAR pattern, consistent with the first phase of the study.

Data Analysis

Item parameters for simulated and PISA 2012 test under CTT and IRT (2-PL) were estimated by using BILOG-MG. Significance of differences between item parameters obtained from complete and real data sets were examined by the Wilcoxon signed rank test and the effect sizes for the significant differences were calculated.

Findings

The findings of the studies made on the simulated data sets are given first and the findings of the studies made on the PISA 2012 data were given afterwards. The results of the Wilcoxon signed ranks test on the comparison of the item parameters obtained by different missing data techniques with the parameters obtained from the complete data sets within the scope of CTT are given in Table 2.

Table 2. Wilcoxon signed ranks test z values of item parameter estimated from simulated data sets according to CTT

Sample Size	Missing Data %	Item Parameters					
		p			d		
		LD	RI	EM	LD	RI	EM
500	%2	-2.97**	.00	.00	-3.08**	+2.83**	+3.00**
	%5	-3.09**	1.73	1.41	-3.07**	+3.17**	+3.11**
	%10	-3.08**	+2.45*	+2.45*	-3.07**	+3.10**	+3.08**
1000	%2	-2.64**	.00	.00	-2.98**	+2.83**	+3.16**
	%5	-3.08**	.00	.00	-3.07**	+3.12**	+3.12**
	%10	-3.07**	.38	.38	-3.06**	+3.10**	+3.13**
2000	%2	-2.46*	.00	.00	-2.96**	+2.24*	+2.24*
	%5	-3.09**	.58	1.00	-3.08**	+3.18**	+3.15**
	%10	-3.07**	1.00	1.00	-3.07**	+3.10**	+3.11**

*p<.05, **p<.01

In Table 2, "-" and "+" symbols in cells where the significant difference is concerned (p<.05 or p<.01) represent the situation of the parameter obtained under the relevant condition according to the complete data set. For example, when the sample size is 500 and the missing data rate is %2, the item difficulties (p) obtained by the listwise deletion (LD) are found to be lower than the complete data sets (z=2.97, p<.01). Similarly, it is seen that the discriminations (d) obtained with EM are higher than all sample sizes and complete data sets for missing data rates included in the study.

Table 2 shows that the item difficulties estimated by the regression imputation and EM do not differ from the complete data sets (n=500 and missing rate=%10, excluded), p>.05; item difficulties estimated by listwise deletion are found to have lower values than complete data sets in all conditions. When the discrimination values are examined, it is seen that the values estimated by listwise deletion are lower in all conditions than the complete data sets, and the values estimated by regression imputation and EM are higher than the complete data sets in all conditions.

The effect sizes suggested by Pallant (2007) were calculated for Wilcoxon signed rank test cases where significant differences were observed and evaluated according to the criterion of effect size >.1 "small",> .3 "medium" and >.5 "large" effect. Since the number of items in each dataset is 12, significant z values obtained from the signed rank test which are greater than 1.47 indicate "medium" and which are greater than 2.45 indicate the "large" effect. In Table 2, when the values for the comparison of item difficulties are examined, it is understood that the effect size is large in all conditions for the listwise deletion. For regression imputation and EM, it is understood that a large effect size is the case only when the sample size is 500 and the missing data rate is %10. Similarly, when the sample size is 2000 and the missing data rate is %2, the effect size is large for all the estimates except for item discriminations obtained from the regression imputation and EM, medium effect. Table 3 shows the χ^2/sd rates for model fit calculated after reaching complete data sets by listwise deletion, regression imputation and EM under 2-PL model.

Table 3. Mean χ^2/sd rates obtained by applying missing data techniques from simulated data sets

Missing Data Technique	N	Missing Data %		
		%2	%5	%10
LD	500	1.91	4.02	10.03
	1000	3.25	7.67	19.82
	2000	8.27	20.00	46.26
RI	500	1.19	1.19	1.29
	1000	1.84	1.90	2.08
	2000	2.63	2.70	2.91
EM	500	1.20	1.19	1.29
	1000	1.83	1.91	2.05
	2000	2.64	2.68	2.91

Table 3 shows that the χ^2/sd rates obtained by the regression imputation and EM are below 3 for all the sample size and missing data rates included in the study. When the χ^2/sd rates obtained from the listwise deletion are examined, it is found that χ^2/sd is smaller than 3 only for the data sets with sample size of 500 and missing data rate of %2; as the sample size or missing data rate increases, χ^2/sd increases up to 46.36, in other words, data - model fit is not achieved.

In order to obtain accurate estimates in the studies conducted within the scope of the IRT, model fit must be satisfactory (Hambleton et al., 1991). Beside this, item parameters have been estimated in order to compare the results obtained using different missing data techniques. The results of the Wilcoxon signed rank tests on the comparison of the item parameters obtained by different missing data techniques with the parameters obtained from the complete data sets are given in Table 4.

Tablo 4. Wilcoxon signed ranks test z values of item parameter estimated from simulated data sets according to IRT

Sample Size	Missing Data %	Item Parameters					
		p			d		
		LD	RI	EM	LD	RI	EM
500	%2	+3.10**	.45	1.00	-3.06**	+2.99**	+3.02**
	%5	+3.07**	-2.11*	1.41	-3.06**	+3.08**	+3.09**
	%10	+3.06**	-2.53*	-2.54*	-3.06**	+3.07**	+3.09**
1000	%2	+2.99**	.33	.82	-2.81**	+3.10**	+3.10**
	%5	+2.94**	.00	.00	-3.06**	+3.08**	+3.08**
	%10	+3.07**	.14	.42	-3.06**	+3.06**	+3.06**
2000	%2	+3.09**	1.41	1.41	-2.94**	+3.10**	+3.10**
	%5	+3.07**	1.41	1.41	-3.06**	+3.07**	+3.07**
	%10	+3.06**	1.73	1.51	-3.06**	+3.06**	+3.06**

*p<.05, **p<.01

Table 4 shows that the b-parameters obtained by the listwise deletion are higher than the complete data sets for all sample sizes and missing data rates included in the study. Lower b-parameters are obtained compared to the complete data sets with regression imputation and EM when the sample size is 500 and the missing data rate is %5 or more. With these two techniques, there is no statistically significant difference between the b parameters obtained in cases where the sample sizes are 1000 and 2000 and the parameters obtained from the complete data sets.

When the results of the comparison of a parameters obtained within the scope of IRT in Table 4 are examined, it is seen that lower a parameters are obtained compared to complete data sets under all conditions with listwise deletion; higher a parameters are obtained compared to complete data sets under all conditions with regression imputation and EM. When the sample size is 500 and the missing data rate is %5, the effect size obtained for the regression imputation technique is medium (.43); for all other situations where the difference is significant, it seems that the effect size is large.

Whether the data of 1482 students who responded to the PM3 subtest in the PISA 2012 study meets the assumption of unidimensionality was examined by means of principal components analysis based on the

tetrachoric correlation matrix. According to the analysis results, Bartlett sphericity test was significant (p <.01) and KMO value was 0.86. When the eigenvalues were examined, it was seen that there were two eigenvalues (5.07 and 1.15) greater than one. The first factor explained 42.26% of the total variance and the factor loadings were found to be between .37 and .81 So, it can be said that the PM3 subtest of 12 items measures one characteristic ability and the assumption of unidimensionality is satisfied. In cases where the assumption of unidimensionality is met, it is assumed that the assumption of local independence is also met (Hambleton et al., 1991). From this point, the item parameters estimated from the complete data sets are given in Table 5.

Table 5. Item parameters of items in PISA 2012 PM3 data

Item	Item Parameters			
	CTT		IRT (χ^2/sd=3,94)	
	p	d	b	a
I1	.49	.39	.07	.48
I2	.41	.56	.37	.84
I3	.37	.50	.60	.67
I4	.68	.57	-.68	.97
I5	.04	.62	2.24	1.32
I6	.58	.47	-.34	.65
I7	.22	.69	1.00	1.30
I8	.58	.49	-.35	.66
I9	.91	.30	-3.20	.47
I10	.22	.66	1.04	1.20
I11	.52	.49	-.08	.67
I12	.28	.48	1.07	.65

Table 5 shows that for the items included in the PM3 subtest, within the scope of CTT, the item difficulties vary between .04 and .91 and the item discrimination vary between .30 and ,69; within the scope of the IRT item difficulties vary between -3.20 and 2.24, while the discriminations vary between .47 and 1.32. The calculated χ^2/sd rate is 3.94. For the data sets obtained by data deletion and missing data techniques at different rates, the satisfaction of the assumption of unidimensionality was investigated by analysis of the principal components based on the tetrachoric correlation matrix. The results of the χ2 / sd values obtained for the principal components analysis and the model - data compatibility are given in Table 6.

Table 6. Results of principal components analysis for PM3 data and χ2 / sd rates

Missing Data Technique	Missing Data %	n	KMO	Bartlett	Eigenvalues	Extracted Variance	Factor Loading		χ^2/sd
							Min.	Max.	
LD	%2	1153	.86	1798.7**	1: 5.20 2: 1.20	43.36%	.38	.82	5.00
	%5	791	.86	1223.5**	1: 5.20 2: 1.25	43.33%	.43	.79	10.67
	%10	429	.83	707.6**	1: 5.21 2: 1.35	43.44%	.51	.81	25.65
RI	%2	1482	.86	2297.1**	1: 5.16 2: 1.17	42.98%	.36	.81	4.00
	%5		.86	2421.6**	1: 5.28 2: 1.18	43.99%	.38	.80	3.61
	%10		.86	2571.5**	1: 5.41 2: 1.15	45.05%	.34	.82	3.71
EM	%2		.86	2294.1**	1: 5.16 2: 1.17	42.97%	.36	.81	3.92
	%5		.86	2413.2**	1: 5.27 2: 1.17	43.95%	.37	.80	3.55
	%10		.86	25915**	1: 5.42 2: 1.14	45.13%	.34	.82	3.66

**p<.01

When principal components analysis result in Table 6 are examined, it can be stated that all data sets meet the unidimensionality assumption. When χ^2/sd rates are examined, it is seen that 2-PL model fit is achieved at data sets completed with regression imputation and EM. When the χ^2/sd rates obtained by the listwise deletion are examined, it is seen that the model fit decreases due to the increase of the missing data rate. This finding is in agreement with the findings in Table 3 for simulated data sets. The results of the Wilcoxon signed rank test on the comparison of item parameters estimated using listwise deletion, regression imputation and expectation - maximization algorithm techniques on PM3 test data within the scope of CTT with the parameters estimated from complete PM3 dataset are given in Table 7.

Table 7. Wilcoxon signed ranks test z values of item parameters estimated from PM3 data according to CTT

Item Parameter	Missing Data Technique	Missing Data %		
		%2	%5	%10
p	LD	-2.94**	-3.06**	-3.06**
	RI	.67	1.29	1.58
	EM	.66	1.38	1.48
d	LD	-2.12**	-3.06**	-3.06**
	RI	-2.83**	-2.94**	-2.82**
	EM	-2.83**	-2.98**	-2.82**

*p<.05, **p<.01

In Table 7, it is seen that item difficulties obtained by listwise deletion within the scope of CTT are lower than the complete data set. There is no significant difference between item difficulties obtained by using regression imputation and EM and item difficulties of complete data set. In Table 7, it is seen that item discriminations obtained by different missing data techniques for all sample size and missing data rates included in the study are lower than the complete data set. The Wilcoxon signed rank test results for comparison of item parameters estimated under the IRT through PM3 data with item parameters obtained from the complete data set are given in Table 8.

Table 8. Wilcoxon signed ranks test z values of item parameter estimated from PM3 data according to IRT

Item Parameter	Missing Data Technique	Missing Data %		
		%2	%5	%10
b	LD	-2.19**	-2.12**	-3.06**
	RI	.20	1.16	1.88
	EM	.13	1.69	-1.96*
a	LD	-2.59**	-3.06**	-3.06**
	RI	-2.76**	-2.51**	-2.82**
	EM	-2.79**	-2.67**	-2.82**

*p<.05, **p<.01

In Table 8, it is seen that item difficulties obtained by listwise deletion are lower than the complete data set. There is no significant difference between item difficulties of complete data set and item difficulties obtained by using regression imputation and EM. It is also seen that the item discriminations obtained by different missing data techniques for all sample size and missing data rates are lower than the complete data set. Comparing Table 2 and Table 7, it is seen that the findings obtained through simulated and real data sets for item difficulties are consistent. According to this, lower item difficulty values are obtained with the listwise deletion than the complete data sets and there is no significant difference between item difficulties obtained from complete data sets and regression imputation and EM (p>.05).

It is seen that there are some inconsistencies for the discrimination values given in Table 2 and Table 7. With the listwise deletion, lower discrimination values are obtained compared to the complete data sets. However, by using regression imputation and EM, higher discrimination values were obtained with simulated data sets and lower discrimination values were obtained from real data sets. A similar situation is seen when Table 4 and Table 8, which contain the results within the scope of IRT, are compared. The results found for the item difficulties estimated by regression imputation and EM are similar. However, with listwise deletion, higher item difficulties were obtained in simulated data sets, and lower item difficulties

were obtained in real data sets. In addition, for regression imputation and EM, higher discrimination values were obtained with simulated data sets, and lower discrimination values were obtained with real data sets.

It is thought that there are two possible sources of this inconsistency; sample size and item characteristics. An additional study was conducted to investigate the difference between the manipulated sample sizes (n=500, n=1000 and n=2000) and the PM3 subtest sample size (n=1482). In this additional study, a new data set of 12 items, which has the same sample size as PM3 test (n=1482) was produced. For this data set, the item parameters estimated within the scope of CTT and IRT are given in Table 9.

Table 9. Item parameters for additional data set (n=1482)

Item No	Item Parameters			
	CTT		IRT(χ2/sd=2,01)	
	p	d	b	a
I1	.20	.77	1.08	1.35
I2	.45	.81	.17	1.53
I3	.28	.71	.80	1.07
I4	.69	.72	-.62	1.49
I5	.20	.86	.98	1.90
I6	.74	.72	-.75	1.76
I7	.40	.60	.41	.78
I8	.31	.76	.65	1.27
I9	.69	.57	-.79	.82
I10	.57	.51	-.33	.64
I11	.38	.88	.34	2.06
I12	.19	.79	1.04	1.44

In Table 9, it is seen that within the scope of CTT, item difficulties obtained is between 0,190 - 0,738 and the discriminations is between 0,514 - 0,881; within the scope of IRT, item difficulties obtained is between -0.791 - 1.084 and the discriminations is between 0,643 - 2,057. The χ2 / sd rate is 2.01. The tests carried out on the PM3 test data are repeated one by one on the additional data set. The results of the Wilcoxon signed rank tests carried out on the additional data set within the scope of CTT are given in Table 10.

Table 10. Wilcoxon signed ranks test z values of item parameters estimated from additional data according to CTT

Item Parameter	Missing Data Technique	Missing Data %		
		%2	%5	%10
P	LD	-3.06**	-3.06**	-3.06**
	RI	.92	.09	.09
	EM	.31	1.74	2.02
D	LD	-2.91**	-3.06**	-3.06**
	RI	-3.06**	-3.06**	-3.06**
	EM	+3.07**	+3.06**	+3.06**

*p<.05, **p<.01

In Table 10, it is seen that item difficulties obtained by listwise deletion within the scope of CTT are lower than the complete data set. And also it is seen that the difference between item difficulties obtained from additional data set and the item difficulties obtained by regression imputation and expectation - maximization algorithm is not significant. When the discrimination values are examined, lower values were obtained with listwise deletion and regression imputation techniques compared with additional data set, and higher values were obtained with EM. The results of the Wilcoxon signed rank test carried out on the additional data set within the scope of IRT are given in Table 11.

In Table 11, it is seen that item difficulties obtained by listwise deletion are higher than the additional data set. It is seen that the difference between item difficulties obtained from additional data set and the item difficulties obtained by regression imputation and EM is not significant. When the discrimination values were examined, lower values were obtained with listwise deletion (excluding the missing data rate of %2)

and regression imputation techniques compared with additional data set, and higher values were obtained with EM.

Table 11. Wilcoxon signed ranks test z values of item parameter estimated from additional data according to IRT

Item Parameter	Missing Data Technique	Missing Data %		
		%2	%5	%10
b	LD	+2.82**	+3.06**	+3.07**
	RI	.43	.63	.98
	EM	1.13	+2.39*	1.65
a	LD	1.84	-2.71**	-3.06**
	RI	-2.80**	-3.06**	-3.06**
	EM	+2.84**	+2.94**	+3.06**

*p<.05, **p<.01

Discussion and Conclusion

In this study, the performances of three different missing data techniques were compared through simulated and real data sets and when the findings for the item parameters within the scope of CTT were examined as a whole, the item difficulties and discriminations obtained with the listwise deletion for the whole sample size and the missing data rates were found to be lower compared with the complete data. It has been observed that the discriminations had higher values when there are no significant differences between the item difficulties obtained by regression imputation and EM and the values obtained from the complete data set (except for a few exceptions). In summary, the listwise deletion causes test items to be more difficult; regression imputation and EM cause them to be more discriminating.

Findings of item difficulties obtained from the PM3 subtest in the PISA 2012 study and from the additional data set included in the survey are in agreement with those obtained in the first stage. Based on these findings, it can be generalized that within the scope of CTT, while lower item difficulties can be achieved by using listwise deletion, item difficulties obtained from complete data can be achieved by using regression imputation and expectation - maximization algorithm. Similarly, with the listwise deletion item discriminations tend to be estimated lower.

Findings that are consistent within the scope of the CTT look contradictory to the findings of Demir (2013). It is known that the performances of the missing data techniques in different missing data patterns can change (Akbaş, 2014). Therefore, it can be assumed that this contradiction results from the fact that the missing data pattern in the data set given in the study by Demir (2013) had MAR; the fact that the missing data pattern included in this study was MCAR.

When the findings obtained from the PM3 subtest and the additional data set are examined, it is seen that while low discrimination values are obtained with the regression imputation technique, inconsistent results are obtained with the EM algorithm. This inconsistency also applies to findings obtained within the scope of IRT. A possible reason for this inconsistency is that the model fit in the simulated data sets is higher. Another reason may be the fact that the parameters of the simulated data change in a narrower range than the real data. This becomes meaningful when evaluated in conjunction with the work of Finch (2008). In this study, it has been determined that the estimates obtained by the EM may vary compared with the parameters obtained from the complete data sets. Accordingly, the fact that the items in the complete data set have low or high discrimination can positively or negatively affect the item parameters obtained from the data sets to be completed with the EM algorithm. When all of the simulated data sets included in this research are considered, b parameters vary between -1.05 and 1.03, and the PM3 test b parameters vary between -3.20 and 2.24 (see Table 5 and Table 9). Similarly, for n = 500, 1000 and 2000 sample sizes, the number of items with a discrimination value greater than one is 5, 5, and 9, respectively, while this number is 3 in the PM3 subtest. Therefore, it can be expressed that the maximization algorithm produces estimations that are difficult to be predicted in data sets with different item parameters. Estimations obtained by the regression imputation technique are consistent in itself for different sample sizes. From this point of view, it can be said that the regression imputation technique is very sensitive to the sample size.

On the other hand, it is understood that the examination of the percentage of missing data to be made over the entire data set is not sufficient. Depending on the sample size of the listwise deletion and the increase in the missing data rate, it has been determined that the model fit which must be achieved within the scope of the IRT is weakened, even exceeding the acceptable limits. Similarly, it has been observed that as sample size increases for the same missing data rate, model fit decreases. So, it is understood that the use of listwise deletion can cause serious problems when the investigations made within the scope of both CTT and IRT are evaluated. It seems that the widespread view that missing data at a ratio of %2 is tolerable, even though it has MCAR pattern, has not been confirmed for both simulated and real data sets. In other words, it can be said that tolerable missing data rate should be below %2. It can be pointed out that the rate of %5 stated by Schafer (1997) and the rate of %10 stated by Hair et al., (2006) are very high. Therefore, in order to be able to analyze data sets containing %2 or more missing data, data collection should be continued until the size of the previously planned sample is reached. In the cases where data collection is not possible, instead of listwise deletion, regression imputation or EM algorithm given in this study, multiple imputation techniques can be preferred, which are determined to be effective even when %10 of data was missing (Doğanay Erdoğan, 2012).

In this study, the cases that the data had MCAR pattern were modeled. In future research, the effectiveness of missing data techniques can be explored through MAR and MNAR data patterns. Considering that the number of missing data techniques and the way in which they are implemented change over time, it can be suggested that similar investigations can be performed using different techniques.

In this study, model fit is examined over χ^2/sd rate within the scope of IRT. It may be considered to repeat the study taking other indicators of fit (G^2, -2LL, standardized residuals, test characteristic curves) into consideration. In addition, within the scope of IRT, there is a need to examine missing data techniques on data sets which have multidimensional models and complex structures.

References

Akbaş, U. (2014). *Farklı örneklem büyüklüklerinde ve kayıp veri örüntülerinde ölçeklerin psikometrik özelliklerinin kayıp veri baş etme teknikleri ile incelenmesi*. Yayınlanmamış doktora tezi, Ankara Üniversitesi, Eğitim Bilimleri Enstitüsü, Ankara.

Allison, P. D. (2002). *Missing data*. Newbury Park, CA: Sage.

Alpar, R. (2011). *Uygulamalı çok değişkenli istatistiksel yöntemler*. Ankara: Detay Yayıncılık.

Bal, C. (2003). *Çok gruplu veri setlerinde eksik gözlem sorununun çözümlenmesi ve sağlık alanında bir uygulama.* Yayınlanmamış doktora tezi, Osmangazi Üniversitesi, Sağlık Bilimleri Enstitüsü, Eskişehir.

Bernaards, C. A. and Sijtsma, K. (1999). Factor analysis of multidimensional polytomous item response data suffering from ignorable item nonresponse, *Multivariate Behavioral Research. 34*(3), 277 – 313. doi:10.1207/S15327906MBR3403_1

Buuren, van S. (2013). *Flexible imputation of missing data*. New York: Chapman & Hall/CRC Press.

Can, S. (2003). *The analyses of secondary education institutions student selection and placement test's verbal section with respect to item response theory models.* Unpublished master thesis, Middle East Technical University, Department of Educational Sciences, Ankara.

Çakıcı Eser, D. (2015). *Çok boyutlu madde tepki kuramının farklı modellerinden çeşitli koşullar altında kestirilen parametrelerin incelenmesi.* Yayınlanmamış doktora tezi, Hacettepe Üniversitesi, Eğitim Bilimleri Enstitüsü, Ankara.

Çelen, Ü. (2008). *Klasik test kuramı ve madde test kuramına dayalı olarak geliştirilen iki testin psikometrik özelliklerinin karşılaştırılması.* Yayınlanmamış doktora tezi, Ankara Üniversitesi, Eğitim Bilimleri Enstitüsü, Ankara.

Çokluk, Ö. and Kayri, M. (2011). Kayıp değerlere yaklaşık değer atama yöntemlerinin ölçme araçlarının geçerlik ve güvenirliği üzerindeki etkisi. *Kuram ve Uygulamada Eğitim Bilimleri*. 11 (1), 289 – 309.

Çüm, S. and 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 iki kategorili puanlanan maddelerden oluşan testlerin psikometrik özelliklerinin incelenmesi*. Yayınlanmamış doktora tezi, Ankara Üniversitesi, Eğitim Bilimleri Enstitüsü, Ankara.

Doğanay Erdoğan, B. (2012). *Çoklu atama yöntemlerinin Rasch modelleri için performansının benzetim çalışması ile incelenmesi*. Yayınlanmamış doktora tezi, Ankara Üniversitesi, Sağlık Bilimleri Enstitüsü, Ankara.

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

Enders, C. K. (2003). Using the expectation maximization algorithm to estimate coefficient alpha for scales with item-level missing data. *Psychological Methods. 8*(3), 322-337. doi:10.1037/1082-989X.8.3.322

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

Finch, H. (2008). Estimation of item response theory parameters in the presence of missing data. *Journal of Educational Measurement. 45(3)*, 225-245.

Field, A. (2005). *Discovering statistics using SPSS*. Sage.

Ginkel, van J. R., Ark, van der L. A. and Sijstma, K. (2007). Multiple imputation for item scores when test data are factorially complex. *British Journal of Mathematical and Statistical Psychology*, 60, 315 - 337. doi:10.1348/000711006X117574

Graham. J. W. (2012). *Missing data analysis and design*. Springer.

Hair, J. F., Black, W. C., Babin, B. J., Anderson, R. E. and Tatham, R. L. (2006). *Multivariate data analysis*. Pearson – Prentice Hall.

Hambleton, R. K., Swaminathan, H. and Rogers, H. (1991). *Fundamentals of item response theory.* Newbury Park CA: Sage.

Han, K. T. (2007). Wingen: Windows software that generates IRT parameters and item responses. *Applied Psychological Measurement*, 31(5), 457-459.

Heerwegh, D. (2005). *Web Surveys. Explaining and Reducing Unit Nonresponse, Item nonresponse and partial nonresponse.* Yayınlanmamış doktora tezi, Katholieke Universiteit Leuven Faculteit Sociale Wetenschappen, Leuven, Belçika.

Holman, R. and Glas, C. A. W. (2005). Modelling non-ignorable missing-data mechanisms with item response theory models. *British Journal of Mathematical and Statistical Psychology*, 58, 1–17. doi:10.1348/000711005X47168

Karasar, N. (2007). *Bilimsel araştırma yöntemi: kavramlar, ilkeler, teknikler*. Ankara: Nobel Yayın Dağıtım.

Köse, İ. A. and Öztemur, B. (2014). Kayıp veri ele alma yöntemlerinin t-testi ve anova parametreleri üzerine etkisinin incelenmesi. *Abant İzzet Baysal Üniversitesi Eğitim Fakültesi Dergisi,* 14(1), 400-412. doi:10.17240/aibuefd.2014.14.1-5000091519

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

Little, R. J. A. (1988). A test of missing completely at random for multivariate data with missing values. *Journal of the American Statistical Association*, 83, 1198-1020. doi: 10.2307/2290157.

Leeuw, E. D. Hox, J. and Huisman, M. (2003). Prevention and treatment of item nonresponse. *Journal of Official Statistics.* 19(2), 153-176.

McKnight, P. E., McKnight, K. M., Sidani, S. and Figueredo, A. J. (2007). *Missing data a gentle introduction*. New York: The Guilford Press.

Nartgün, Z. (2015). Comparison of various methods used in solving missing data problems in terms of psychometric features of scales and measurement results under different missing data conditions. *International Online Journal of Educational Sciences, 7*(4), 252 – 265. doi: 10.15345/iojes.2015.04.017

OECD. (2014). *PISA technical report*. OECD Publishing.

Özer Özkan, Y. (2012). *Öğrenci başarılarının belirlenmesi (öbbs) sınavından klasik test kuramı, tek boyutlu ve çok boyutlu madde tepki kuramı modelleri ile kestirilen başarı puanlarının karşılaştırılması.* Yayınlanmamış doktora tezi, Ankara Üniversitesi, Eğitim Bilimleri Enstitüsü, Ankara.

Roth, P. L. (1994). Missing data: A conceptual review for applied psychologists. *Personnel Psychology,* 47(3), 537-560. doi: 10.1111/j.1744-6570.1994.tb01736.x

Rubin, D. B. (1976). Inference and missing data. *Biometrika 63*, 581-592. doi:10.1093/biomet/63.3.581

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

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

Sijtsma, K. and Ark, van der, L. A. (2003). Investigation and treatment of missing item scores in test and questionnaire data. *Multivariate Behavioral Research. 38*(4), 505-528. doi:10.1207/s15327906mbr3804_4

SPSS (2007). *SPSS missing values* 17.0. SPSS Inc.

Şahin Kürşad, M. (2014). *Sıklıkla kullanılan kayıp veri yöntemlerinin betimsel istatistik güvenirlik ve geçerlik açısından karşılaştırılması.* Yayınlanmamış yüksek lisans tezi, Abant İzzet Baysal Üniversitesi, Eğitim Bilimleri Enstitüsü, Bolu.

Tabachnick, B. G. and Fidell, L. S. (1996). *Using multivariate statistics.* Harper Collins College Publishers.