Missing Value Effect on the Estimation of Item Response Theory Parameters: A Simulation Study by Increasing the Item Difficulty Level

Alejandra Arias Salazar*
University of Costa Rica, San José, Costa Rica – alejandra.ariassalazar@ucr.ac.cr

Abstract

Missing values are frequently found in real data sets. In some instrument designs, for example, in measuring academic achievement, the item difficulty level increases. Therefore, it is common to get unfilled items at the end of the test. The main purpose of this paper is to explore the effect of missing values percentages (10, 20 and 30%) in parameter estimation for a two parameter logistic Item Response Theory model, when the difficulty in the test is ascendant (from -2 to 2) and the discrimination parameter is set randomly. Moreover, we seek to determine the best way to treat those missing values using two imputation methods: treat them as incorrect answers and by mean substitution. Through a simulation study (n=1000, 45 items and 100 replications), we found that in both imputation methods, the difficulty parameter increases as the percentage of missingness increases. The pattern is not clear enough in the discrimination parameter for both imputation methods. Furthermore, if there are 20% of unfilled items, the best way to treat them is by mean substitution (due to a lower Root Mean Square Error and a lower bias); but, if this value increases to 30%, replacement by zero seems to be the best option.

Keywords: Root Mean Square Error; bias; discrimination parameter.

Introduction

Missing data are frequently found in many different kinds of data sets, for example, the health registry, social surveys, educational assessments or psychometrical evaluations. According to previous studies, there are several reasons why incomplete information is generated. On the one hand, it may occur because subjects do not answer one or more items, ignore the answer, have not enough time to finish it, skip it unintentionally, or are afraid to guess. On the other hand, it is possible to get non-responses due to a poor collection and data entry methods, or an intentional test design (Osborne, J. 2013; Finch, H. 2008, DeMars, C. 2002).

Many researches indicate that there are three types of missing data: *missing at random* (MAR), if the probability of missing depends only on the value of other variables (Magnani, M. 2004); *missing completely at random* (MCAR), when there is no systematic cause; and *missing not at random* (MNAR), if there is a relationship between the missingness in a specific variable and the variable itself (Finch, H. 2009). In this study, we are focused on the last case.

It is necessary to treat those missing values in order to conduct a proper analysis since standard methods are not applicable with unanswered items (Pimentel, J. 2005); and also, dealing with the specification of a full parametric model is an even more complex task (Schafer, J. and Graham, J. 2002). Because of this, choosing a correct method to handle them is essential.

Pimentel J. (2005), mentioned common methods to deal with missing information: deleting those cases with at least one item unanswered, employing an imputation technique to replace them, ignoring them for the analysis (not the whole case but the specific unfilled item), and through modeling methods to incorporate missing data.

Frequently, in educational tests (for example, knowledge, capability and skills test to measure academic achievement), the difficulty level of items increases. An example of this case is PISA evaluations (OECD, n.d). When a subject do not answer some questions, those items are usually treated as incorrect answers; in other words, examinees get zero score for each item unfilled. (Finch, H. 2008).

However, Lord (1994) point out that omitted items cannot be treated as incorrect answers to estimate ability and item parameters when there is a time limit to complete the instrument. The reason is that this would lead us to a wrong interpretation about ability information.

It is very important to highlight that time is an essential factor in this kind of tests. That means that the ability of the subject to complete the questionnaire successfully is taken into account, but the subject' speed as well.

In this research, we want to study the effect of missing values on difficulty and discrimination parameters considering a pattern in which most difficult items are located at the end of the instrument. Then, we want to evaluate the consequences of replacing missing values using incorrect answers or mean substitution. However, it is necessary to assume that the examinee follows an order to complete the questionnaire items.

To develop this study, we applied a two parameter logistic model (2PL). The advantage of choosing this model is to establish a non-linear relationship between latent traits and subjects abilities (López, J. 1995). The dichotomous 2PL is mathematically specified as:

$$P\left(Y_{ij} = 1 \middle| \theta_i\right) = \frac{1}{1 + \exp\left[-1.7\, a_j\left(\theta_i - b_j\right)\right]}$$

where $P\left(Y_{ij} = 1 \middle| \theta_i\right)$ is the probability that an examinee i with an ability θ_i correctly answer an item j , a_j indicates the item discrimination parameter, and b_j is the item difficulty parameter.

Methods

The simulation study design includes a sample size of 1000, 45 items and 100 replications. The percentage of missing values was fixed based on two studies. For example, Rubright, J., Nandakumar, R. & Glutting, J. (2014) in *A Simulation Study of Missing Data with Multiple Missing X's* used three levels of unfilled data: 10, 25 and 50%. In a similar way, Holmes Finch (2008) in *Estimation of Item Response Theory Parameters in the Presence of Missing Data* used 5, 15 and 30% of missing values.

To take into account studies already mentioned, we used the last 10 items (of 45) to replace the 10, 20 and 30% with missing values, with the purpose of studying their impact in the parameter estimation (difficulty and discrimination).

We decided to work with 100 replications because Finch H. (2008), Pimentel J. (2005) and Lopez J. (1995) have argued that 100 replications is an appropriate number to study the parameters behavior when there are missing values. Difficulty of items were set from -2 to 2, where this level increase each 5 items (as is shown in table 1). The discrimination parameter was set randomly from a normal distribution with mean zero and standard deviation of 1.

Table 1. ***Item Difficulty level***

Item	Difficulty level
1-5	-2
6-10	-1.5
11-15	-1
16-20	-0.5
21-25	0
26-30	0.5
31-35	1
36-40	1.5
41-45	2

Later, those parameters were analyzed using two imputation methods through different percentages of missing values (already mentioned): treat them as incorrect answers (IN), because this is the common way to deal with them and replace them by mean (ME), as an alternative option.

Finally, to assess the performance of each imputation method, we compared the average values for relative bias and for root mean square error (RMSE) across replications for each parameter estimation. The data generation, analysis and parameter estimation were carried out in R, version 3.1.2 (R Core Team, 2014).

Results

Missing value replaced by zero (incorrect answer)

Table 2 shows parameters of difficulty and discrimination at different levels of missingness replaced by zero. It is possible to identify an increase in the difficulty parameter value from estimated value (without missingness) to the parameter with 10% of unfilled answers. The same increasing pattern occurs from the parameter with 10% of missing values to 20%, except for question 43 because this value decreases from 1.96 to 1.92, and in item 42, the parameter value stays the same. In the case of the difficulty parameter by 30% of omitted responses, each value is higher than previous estimations.

Table 2. ***Difficulty and discrimination parameter estimations by 10, 20 and 30% of missing values replaced as incorrect answers (IN)***

	Difficulty Parameter				**Discrimination Parameter**			
Item	**Estimated**	**Missing value**			**Estimated**	**Missing value**		
		10%	**20%**	**30%**		**10%**	**20%**	**30%**
36	1.44	1.83	2.33	3.30	0.45	0.40	0.41	0.31
37	1.43	1.55	1.62	1.73	2.50	2.32	2.28	2.11
38	0.61	1.61	1.87	3.44	0.29	0.25	0.27	0.19
39	1.53	2.18	2.26	3.97	0.39	0.34	0.42	0.26
40	1.57	1.68	1.80	1.96	1.40	1.35	1.39	1.27
41	2.54	3.43	3.50	5.85	0.15	0.16	0.19	0.14
42	1.96	2.13	2.13	2.30	2.29	2.09	2.07	2.04
43	1.79	1.96	1.92	2.14	1.94	1.76	1.92	1.70
44	2.22	2.39	2.69	2.56	0.75	0.75	0.70	0.78
45	2.08	2.42	2.45	2.48	0.91	0.83	0.84	0.90

Moreover, there is not a clear pattern in the discrimination parameter in comparison with the difficulty parameter. Discrimination decreases from the estimated value to the parameter at 10% of incomplete items, but the relationship between parameters at 10% and 20% is not clear; we observe an increase in discrimination for some items but a decrease in others. Estimations with 30% of missing values are the lowest (except for item 44).

Missing value replaced by mean (ME)

Regarding the difficulty parameter (table 3), there is a very clear ascending pattern. Estimated values are the lowest ones, and difficulty increases as the percentage of missing value increases. The same as replacement by zero method, items 42 and 43 showed a different behavior.

Additionally, not in every case discrimination decreases when the percentage of the missing value is ascendant. However, with 30% of unanswered items, discrimination is lower than the parameter estimate value (without missingness).

Table 3. ***Difficulty and ability parameter estimations by 10, 20 and 30% of missing values replaced by mean (ME)***

	Difficulty Parameter				**Discrimination Parameter**			
Item	**Estimated**	**Missing value**			**Estimated**	**Missing value**		
		10%	**20%**	**30%**		**10%**	**20%**	**30%**
36	1.44	1.94	2.12	2.37	0.45	0.39	0.44	0.43
37	1.43	1.50	1.64	1.68	2.50	2.50	2.14	2.30
38	0.61	1.25	1.67	3.76	0.29	0.27	0.28	0.18
39	1.53	2.10	2.53	3.44	0.39	0.37	0.35	0.29
40	1.57	1.74	1.69	1.86	1.40	1.32	1.45	1.32
41	2.54	3.56	5.43	7.44	0.15	0.14	0.13	0.13
42	1.96	2.10	2.10	2.09	2.29	2.15	2.26	2.17
43	1.79	1.91	2.14	2.11	1.94	1.79	1.62	1.84
44	2.22	2.34	2.56	3.26	0.75	0.75	0.72	0.64
45	2.08	2.39	2.31	2.51	0.91	0.84	0.90	0.89

Assessment of imputation methods

We compared the average values for root mean square error (RMSE) and for relative bias across replications for each parameter estimation. Figure 1 shows that we got the RMSE for both imputation methods at 10% of missing values. However, when this percentage increases to 20%, RMSE is large if we replace those values as incorrect answers, but this scenario changes if the quantity of unfilled items increase to 30%, where imputation by mean got the highest RMSE.

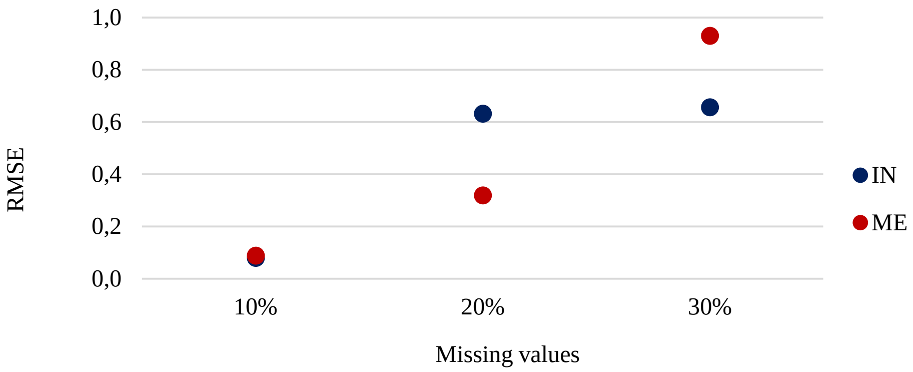

Figure 1. *Average of Root Mean Square Error by level of missing value*

As can be seen in figure 2, the bias average is almost the same at 10% of missing values, both in the IN replacement and in the ME replacement method. At 20% of missing values, we got a high relative bias for IN substitution; but, when the percentage increases to 30%, this bias level is the highest, although the difference is relatively small (0.34 for IN imputation and 0.35 for ME method).

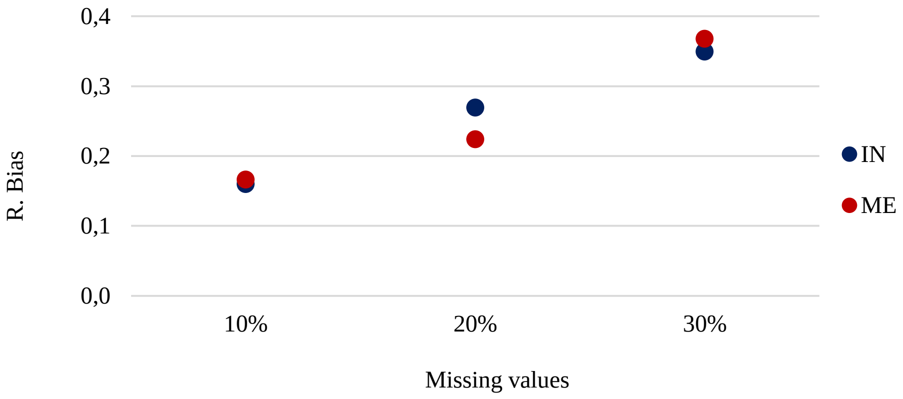

Figure 2. *Average of bias by level of missing value*

Conclusions

Based on the results obtained by the 2PL model with 10, 20 and 30% of missing data, we found that in both imputation methods (replacement by zero and by mean), the difficulty parameter increases as the percentage of missingness increases.

Furthermore, pattern is not clear enough in the discrimination parameter for both imputation methods. However, in almost every item, we found the lowest ability value when there is a bigger percentage (30%) of unfilled items.

As expected, RMSE increases when the percentage of missing values increases. Nevertheless, we notice that in the case of substitution by zero, there is not a big difference between the RMSE at 20% of missingness and the RMSE at 30%. The opposite case is the imputation method by mean, because the RMSE increases substantially if there are 30% of unfilled items. Regarding relative bias assessment, both imputation methods showed a similar behavior.

Lord (1974) and Finch (2008) mentioned that, to treat missing values as incorrect answers is not an optimal method to estimate ability or difficulty parameter because it produces a higher bias than other imputation methods, and also, difficulty is usually overestimated. In spite of this, it is the usual method. In this study we found that, taken into account a test design where the difficulty of items increases, it is better to treat missing values by mean substitution if the percentage of unfilled items is around 20%, but when this percentage increases (to 30%), this method is not the best option since RMSE and bias increases, therefore is better treat them as incorrect answers.

References

DeMars, C. (2002). Missing Data and IRT Item Parameter Estimation. Meeting of the American Educational Research Association. Chicago.

Finch, H. (2008). Estimation of Item Response Theory Parameters in the Presence of Missing Data. Journal of Educational Measurement, 45(3), 225-245.

López, J. (1995). Estimación de parámetros en la TRI: una evaluación BILOG en muestras pequeñas. Psicothema, 7(1), 174-185.

Lord, F. M. (1974). Estimation of latent ability and item parameters when there are omitted responses. Psychometrika, 39(2), 247-264.

Magnani, M. (2004). Techniques for dealing with missing data in knowledge discovery tasks 15(01). Retrieved December 18, 2014, from http://magnanim. web. cs. unibo. it/index. html

Organisation for Economic Co-operation and Development (n.d.) PISA FAQ, Understanting the results, para. 2. Retrieved January 7, 2015, from: http://www.oecd.org/pisa/aboutpisa/pisafaq.htm

Osborne, J. (2013) Dealing with Missing or Incomplete Data: Debunking the Myth of Emptiness. In Best Practices in Data Cleaning: A Complete Guide to Everything You Need to Do Before and After Collecting Your Data (pp. 105-138). SAGE.

Pimentel, J. (2005). Item response theory modeling with non- ignorable missing data (Doctoral Thesis. University of Twente, The Netherlands

R Core Team (2014). R: A language and environment for statistical computing. R Foundation for Stati stical Computing, Vienna, Austria. URL http://www.R-project.org/

Rubright, J. D., Nandakumar, R., & Glutting, J. J. (2014) A Simulation Study of Missing Data with Multiple Missing X's. Practical Assessment, Research & Evaluation, 19(10). Retrieved December 16, 2014, from http://pareonline.net/

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